

# Model-Based Anti-HIV Therapy

## **Dissertation**

zur Erlangung des Grades  
des Doktors der Naturwissenschaften (Dr. rer. nat.)  
der Naturwissenschaftlich–Technischen Fakultäten  
der Universität des Saarlandes  
von

**Tobias Sing**

Saarbrücken  
2007

Tag des Kolloquiums: 17. 4. 2008

Dekan: Prof. Dr. Thorsten Herfet

Berichterstatter: Prof. Dr. Thomas Lengauer, Ph.D.  
Prof. Dr. Hans-Peter Lenhof  
Prof. Dr. Joachim Selbig

Prüfungsausschuss: Prof. Dr. Manfred Pinkal (Vorsitzender)  
Prof. Dr. Thomas Lengauer, Ph.D.  
Prof. Dr. Hans-Peter Lenhof  
Prof. Dr. Joachim Selbig  
Dr. Francisco Domingues (Akademischer Beisitzer)





Gerda Sing (1943–2005)

To the Memory of my Mother,  
in Love and Gratitude.

## Abstract

Modern combination drug therapy has substantially improved the clinical management of HIV-1 infection. Still, the emergence of drug-resistant variants eventually leads to therapy failure in most patients. The selection of an optimal follow-up regimen is complicated by an ever-increasing range of possible drug combinations. In this thesis, we present foundations for rational, model-based treatment strategies.

Firstly, we study viral evolution. In a simulation study, we establish a general link between the shape of a fitness landscape and population dynamics, using an idealized population undergoing mutation, recombination and selection at three biallelic loci as an example. Using techniques from survival analysis, a model of mutation dynamics in the absence of drug is proposed. Differently from mutation accumulation, mutations are found to disappear independently from each other, but with individual survival probabilities. A Fisher kernel for mixtures of mutagenetic trees is derived, quantifying the similarity of evolutionary escape from drug pressure between two viral sequence samples. Kernel-based prediction of drug resistance leads to significant improvements over an evolution-agnostic approach.

Secondly, the controversial interplay between genotypic, phenotypic, and clinical resistance is analyzed. Methods for identifying resistance mutations from either *in vitro* or *in vivo* data, and for characterizing mutational covariation patterns are described. A case study focusing on reverse transcriptase inhibitors yields over 20 previously undescribed mutations, most of them extending classical resistance pathways. Finally, the widely held view of an ill-defined relation between phenotypic and clinical resistance is challenged, and a hybrid model incorporating both genotypic and inferred phenotypic information is shown to outperform its components in predictivity. Surprisingly, the incorporation of viral fitness does not lead to any further improvements.

Thirdly, genotypic, clinical, host-specific, and structural determinants of viral cell entry via the human receptors CCR5 and CXCR4 are investigated. Previously undescribed mutations in the third hypervariable region of the viral envelope protein, CD4<sup>+</sup> cell counts, host heterozygosity for the CCR5- $\Delta$ 32 allele, number of sequence ambiguities arising from population sequencing, and presence of indels are shown to be predictive of coreceptor choice alone and in combination, as are changes in the predicted side-chain conformation in a three-dimensional model of the V3 loop.

## Kurzfassung

Moderne Kombinationstherapien haben die klinische Behandlung von HIV-1-Infektionen wesentlich verbessert. Trotzdem führt das Auftauchen wirkstoffresistenter Varianten bei den meisten Patienten letztlich zum Versagen der Therapie. Die Auswahl eines optimalen Nachfolgeregimes wird durch die stetig wachsende Zahl möglicher Wirkstoffkombinationen erschwert. In dieser Arbeit präsentieren wir Grundlagen für rationale, modellbasierte Behandlungsstrategien.

Als erstes untersuchen wir die virale Evolution. In einer Simulationsstudie etablieren wir eine allgemeine Verbindung zwischen der Form einer Fitnesslandschaft und der sich ergebenden Populationsdynamik, am Beispiel einer idealisierten Population unter dem Einfluss von Mutation, Rekombination und Selektion an drei biallelischen Loci. Unter Verwendung von Techniken der Ereigniszeitanalyse wird ein Modell zur Mutationsdynamik in Abwesenheit von Wirkstoffen vorgeschlagen. Anders als bei der Akkumulation von Mutationen geschieht deren Verschwinden unabhängig voneinander, jedoch mit individuellen Überlebenswahrscheinlichkeiten. Ein Fisher-Kern für Mixturen mutagenetischer Bäume wird hergeleitet. Dieser quantifiziert die Ähnlichkeit der evolutionären Flucht vor Medikamentendruck zweier viraler Sequenzen. Die kernbasierte Vorhersage von Wirkstoffresistenz führt zu einer signifikanten Verbesserung verglichen mit einer evolutions-agnostischen Methode.

Als zweites analysieren wir das kontroverse Zusammenspiel von genotypischer, phänotypischer und klinischer Resistenz. Wir beschreiben Methoden zur Identifikation von Resistenzmutation auf der Basis von *in vitro* oder *in vivo*-Daten, und zur Charakterisierung von Kovariationsmustern zwischen Mutationen. Eine Fallstudie mit Fokus auf Inhibitoren der Reversen Transkriptase ergibt über 20 bisher unbeschriebene Mutationen; die meisten von ihnen erweitern klassische Resistenzpfade. Schließlich widerlegen wir die häufig geäußerten Zweifel an einem deutlichen Bezug zwischen phänotypischer und klinischer Resistenz. Ein hybrides Modell, das sowohl auf genotypischer als auch auf inferierter phänotypischer Information beruht, übertrifft seine Bestandteile an Prädiktivität. Überraschenderweise führt das Einbeziehen von viraler Fitness zu keinen weiteren Verbesserungen.

Als drittes studieren wir genotypische, klinische, wirtsspezifische und strukturelle Determinanten des viralen Zelleintritts über die menschlichen Rezeptoren CCR5 und CXCR4. Bislang unbeschriebene Mutationen in der dritten hypervariablen Region des viralen Hüllproteins, die Anzahl an CD4<sup>+</sup>-Zellen, die Heterozygotie des Wirts für das CCR5- $\Delta$ 32 Allel, die Anzahl der sich durch Populationssequenzierung ergebenden Sequenzambiguitäten, und das Vorhandensein von Indels stellen sich als prädiktiv für die Korezeptorwahl heraus, sowohl alleine als auch in Kombination. Dasselbe gilt für Änderungen in der vorhergesagten Seitenkettenkonformation in einem dreidimensionalen Modell der V3-Schleife.

## Acknowledgements

This work was carried out in the *Department for Computational Biology and Applied Algorithmics* at the *Max Planck Institute for Informatics* in Saarbrücken. First and foremost, I would like to thank my advisor Thomas Lengauer for his support and advice, and for providing me with all the freedom to follow own ideas. I stand on the shoulders of Niko Beerenwinkel, whose Ph.D. thesis ([Beerenwinkel, 2004](#)) and other works have strongly influenced the research presented here. Niko, thanks for introducing me to the field, for all the joint work, and for the time with you in Berkeley.

Thanks also to all former and present members of our group for the unique working atmosphere, the many helpful discussions, and the fun activities besides work. Special thanks go to Oliver Sander for his friendship, countless discussions, and the joint work on ROCr and HIV coreceptor usage. Thanks to Jörg Rahnenführer, Francisco Domingues, and Christoph Bock for many discussions of statistical and biological issues. Thanks to my fellow graduate students in HIV bioinformatics for being such a good team: Andre Altmann, Jochen Maydt, Kirsten Roomp, and Alexander Thielen. Thanks to the students who worked with me on a number of projects: Jan Christoph, Peter Schüffler, Dominik Selzer, and Hendrik Weisser. Thanks to Achim Büch and Ruth Schnepfen-Christmann who both contributed a lot to making our group such a nice place.

This work would not have been possible without wonderful collaborators: Francesca-Ceccherini Silberstein, Valentina Svirch, Carlo-Federico Perno (University of Rome “Tor Vergata”); Richard Harrigan, Andrew Low (British Columbia Center for Excellence in HIV/AIDS); Herman Van Vlijmen, Hans Vermeiren, Elke Van Craenenbroeck, Bart Winters, Koen Van der Borght (Virco BVBA, Mechelen); Lee Bachelier (VircoLab, Inc., Durham, NC); the members of the *Genafor* group, especially Martin Däumer, Rolf Kaiser (University of Cologne), Hauke Walter, Monika Tschochner, Klaus Korn (University of Erlangen); Bernd Kupfer (University of Bonn); the members of the *EuResist* project, especially Mattia Prospero (University of Rome) and the coordinators Maurizio Zazzi (University of Siena), Francesca Incardona (Informa, Rome); Annemie Vandamme, Koen Deforche (Rega Institute, University of Leuven); Bob Shafer, Soo-Yon Rhee (Stanford University).

I would like to thank the German Academic Exchange Service for funding my stay in Berkeley (grant D/06/41866), and the German National Academic Foundation for their support during both my undergraduate and graduate studies.

Thanks to the developers of free software, especially those involved in the R Project for Statistical Computing.

Last but not least I would like to thank my family and friends for their love and support, especially Anne, my father, Claudia and Sebastian, my grandmother — and my grandfather and mother in Heaven.

---

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Anti-HIV Therapy</b>	<b>9</b>
2.1	Epidemiology . . . . .	9
2.2	Structure of virion, genome, and viral proteins . . . . .	11
2.3	Replication and pathogenesis . . . . .	15
2.4	Anti-HIV drugs . . . . .	18
2.5	Evolution, drug resistance, combination therapy . . . . .	21
2.6	Challenges and opportunities for modeling . . . . .	24
<b>3</b>	<b>Modeling: The Views of Aristotle and Plato, and Later Developments</b>	<b>27</b>
3.1	Generality, predictive power, explanatory power . . . . .	27
3.2	Scientific realism vs. instrumentalism . . . . .	28
3.3	Explanatory vs. empirical models . . . . .	29
3.4	Statistical vs. “machine learning” models . . . . .	30
3.5	Examples . . . . .	34
3.5.1	Empirical models and the “Rashomon” effect . . . . .	34
3.5.2	Modeling planetary motion . . . . .	34
3.5.3	Modeling sunflower growth . . . . .	37
3.5.4	Modeling in population genetics . . . . .	38
3.5.5	Summary . . . . .	40
3.6	Modeling strategies in this thesis . . . . .	43
<b>4</b>	<b>Viral Evolution: Foundations and Clinical Implications</b>	<b>47</b>
4.1	Shapes of fitness landscapes and population dynamics . . . . .	48

---

4.1.1	Introduction . . . . .	48
4.1.2	The shape of a fitness landscape . . . . .	49
4.1.3	Evolutionary dynamics . . . . .	51
4.1.4	Symmetries of the ODE system . . . . .	53
4.1.5	Simulations . . . . .	54
4.1.6	Comparison with a rank ordering classification . . . . .	57
4.1.7	Discussion . . . . .	60
4.2	Mutation dynamics during interruption of antiretroviral therapy . . . . .	61
4.2.1	Introduction . . . . .	61
4.2.2	Materials and Methods . . . . .	63
4.2.3	Results . . . . .	65
4.2.4	Discussion . . . . .	73
4.3	Improved prediction of drug resistance by evolutionary modeling . . . . .	75
4.3.1	Introduction . . . . .	75
4.3.2	Background: Mixture models of mutagenetic trees . . . . .	77
4.3.3	The MTreeMix Fisher kernel . . . . .	79
4.3.4	Experimental results . . . . .	82
4.3.5	Conclusions . . . . .	84
4.3.6	Acknowledgment . . . . .	84
<b>5</b>	<b>Drug Resistance: Genotype, Phenotype, and Virological Response</b>	<b>87</b>
5.1	Algorithms for the interpretation of drug resistance: a review . . . . .	87
5.1.1	Introduction: selection of antiretroviral therapy . . . . .	87
5.1.2	The factors of interest . . . . .	88
5.1.3	Predicting <i>in vivo</i> antiviral activity of individual drugs . . . . .	90
5.1.4	Predicting <i>in vivo</i> drug activity from the viral phenotype . . . . .	93
5.1.5	Predicting <i>in vitro</i> resistance and replication capacity from the viral genotype . . . . .	93
5.1.6	Predicting <i>in vivo</i> virological response to combination therapy . . . . .	95
5.1.7	Summary . . . . .	97
5.1.8	The future . . . . .	97
5.2	Discovery and characterization of novel HIV drug resistance mutations . . . . .	99
5.2.1	Introduction . . . . .	100
5.2.2	Mining for novel mutations . . . . .	101
5.2.3	Identifying mutational clusters . . . . .	101

---

5.2.4	Phenotypic characterization of novel mutations . . . . .	105
5.2.5	Discussion . . . . .	109
5.3	Inferring response from genotype: with or without predicted phenotypes?111	
5.3.1	Introduction . . . . .	111
5.3.2	Materials and Methods . . . . .	114
5.3.3	Results . . . . .	119
5.3.4	Discussion . . . . .	125
5.4	Inferring response from genotype: with or without predicted replica- tion capacity? . . . . .	126
5.4.1	Introduction . . . . .	126
5.4.2	Materials and Methods . . . . .	127
5.4.3	Results . . . . .	128
5.4.4	Discussion . . . . .	129
<b>6</b>	<b>Genotypic, Clinical, and Structural Predictors of HIV Coreceptor Usage</b>	<b>133</b>
6.1	Introduction . . . . .	133
6.2	Predicting HIV coreceptor usage from sequence . . . . .	136
6.2.1	Related work . . . . .	136
6.2.2	Method comparison on clonal data . . . . .	138
6.2.3	Method comparison on population-based data . . . . .	141
6.3	Clinical and host markers in predicting coreceptor usage . . . . .	143
6.4	Structural determinants of coreceptor usage . . . . .	150
6.4.1	The structural basis of coreceptor usage . . . . .	150
6.4.2	Methods . . . . .	150
6.4.3	Results . . . . .	154
6.4.4	Discussion . . . . .	156
	<b>Summary</b>	<b>159</b>
	<b>Zusammenfassung</b>	<b>163</b>
	<b>Bibliography</b>	<b>166</b>
	<b>Appendix: List of Publications</b>	<b>195</b>



---

# Introduction

Model-based anti-HIV therapy refers to treatment decisions made on the basis of quantitative models. Ideally, these models should be

- predictive (in close agreement with observed and previously unseen data)
- explanatory (formulated as to reflect the “true” nature of the process to be modeled)
- general (covering all aspects in the interactions between host, virus, and drug)

Such models have a century-old tradition in physics, a younger tradition in chemistry, and are now becoming increasingly popular in the life and medical sciences ([Cohen, 2004](#)).

Modern statistical learning methods allow for deriving predictive models from data when relatively little prior knowledge is available. Thus, they present a powerful starting point into new research fields. Starting from such agnostic, empirically derived models, a unifying theme of this thesis lies in various attempts to move towards more predictive, explanatory and/or general models. These attempts are based on the incorporation of problem-specific background knowledge or on the interconnection of different pieces of information.

The remainder of this introductory chapter will set the stage by discussing the scope of the AIDS pandemic, the basic principles of anti-HIV therapy, the past and future role of HIV as a test for model-best methodologies, and the structure of this thesis.

## HIV/AIDS: scope of the problem

Since its recognition in 1981, AIDS has killed more than 25 million people, making it one of the most destructive epidemics in the history of mankind. Recent years have seen promising developments, including increased access to treatment and prevention programmes. Still, the number of people living with HIV continues to grow, as does the

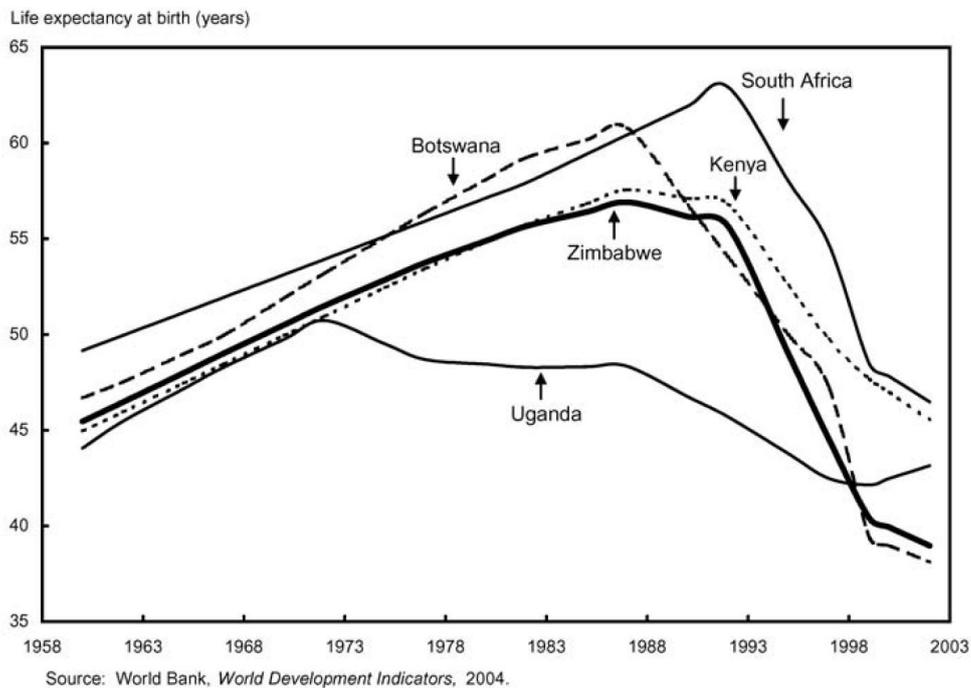


Figure 1.1: Life expectancy at birth in selected African countries, from (United States, 2005). By 1999, HIV/AIDS had cut life expectancy in Africa by one quarter (Logie, 1999). In the same time, life expectancy continued to rise monotonically in all other parts of the world (Dorling et al., 2006). It has been estimated that six years of the difference in life expectancy between Africa and North America is accounted for by differences in access to antiretroviral therapy and means of prevention (Dorling et al., 2006).

number of deaths due to AIDS. In 2006, a total of 39.5 million people were living with HIV, representing an increase of 2.6 million since 2004. Among the adult population (15–49 years), the global prevalence of HIV-infected people is estimated to be 1%. In 2006, an estimated 4.3 million of adults and children were newly infected with HIV, which is about 400,000 more than in 2004. An estimated 2.9 million of adults and children died due to AIDS during the last year (UNAIDS, 2006). Annual HIV/AIDS deaths have been projected to rise to 6.5 million by 2030, under the assumption that coverage with antiretroviral drugs reaches 80% by 2012 (Mathers and Loncar, 2006). In this scenario, the total number of deaths from HIV/AIDS between 2006 and 2030 is projected to be 117 million. The impact of the AIDS pandemic is so severe that it is addressed in one of the eight UN Millennium Development Goals, specifically “by 2015, to halt and begin to reverse the spread of HIV/AIDS”. (Annan, 2000, ch. 7)

This thesis is concerned with the development of rational tools for decision support in anti-HIV therapy. However, it must not be forgotten that in many developing

countries, less than one quarter of those in need are actually receiving antiretroviral therapy (UNAIDS, 2006). This is in stark contrast to the idealistic formulation of the Constitution of the World Health Organization, which states that “the highest attainable standard of health is one of the fundamental rights of every human being without distinction of race, religion, political belief, economic or social conditions” (World Health Organization, 2005, p.1). Further discussion of this issue is beyond the scope of this thesis, but Figure 1.1 will serve as a dramatic reminder of the limited impact of scientific contributions on public health in the absence of sufficiently effective political and economical measures.

### Principles of anti-HIV therapy

Patients infected with the virus HIV-1 exhibit a gradual decline in their naive and memory CD4<sup>+</sup> T-lymphocyte populations (Douek et al., 2003). Due to their coordinating role, depletion of CD4<sup>+</sup> cells disrupts many defense mechanisms of the immune system. Patients infected with HIV eventually die from diseases rarely observed or leading only to mild symptoms in HIV-negative individuals, including specific infections, parasitic diseases (Bonnet et al., 2005), or malignant cancers (Bonnet et al., 2004). Given the central role of CD4<sup>+</sup> cell destruction in the morbidity and mortality of HIV-infected patients, it is understandable that current treatment guidelines (DHHS Panel on Antiretroviral Guidelines for Adults and Adolescents, 2006) consider the ultimate goals of anti-HIV therapy “to prevent further immune deterioration” (immunological goal), and “to avoid HIV-associated morbidity and mortality” (clinical goal). The AIDS disease is the ultimate consequence of treatment that does not meet these goals. Per definition, AIDS is diagnosed in HIV-positive individuals with CD4<sup>+</sup> cell count below 200 cells/mm<sup>3</sup>, or CD4 percentage below 14%, or at least one from a list of currently 26 “AIDS-defining” diseases (CDC, 1992).

If CD4<sup>+</sup> cells are killed at a certain rate, stable CD4<sup>+</sup> cell counts (and thus fulfillment of the ultimate goals of therapy) could be obtained by either increasing the rate of production of novel CD4<sup>+</sup> cells, or by decreasing their rate of destruction. Increasing the production rate is being investigated in the field of immunotherapy, but the clinical efficacy of these experimental treatments remains to be proven (Wynne and Davey, 2005). While immunotherapy is targeted at relieving the symptoms of CD4<sup>+</sup> cell decline by ensuring enough replenishment of cells, decreasing the rate of destruction requires targeting the cause(s) underlying this process. Currently, it is hypothesized that the depletion is mostly attributable to the direct effects of HIV replication in this cell population (however, cf. Section 2.3). Specifically, HIV RNA plasma level has been considered the main determinant of the speed of CD4<sup>+</sup> cell loss. Based on this hypothesis, “suppression of viremia to less than detection limits” (DHHS Panel on Antiretroviral Guidelines for Adults and Adolescents, 2006) is recommended as the direct (virological) goal of anti-HIV therapy.

To date, 21 compounds from four different classes of drugs have been approved for anti-HIV therapy by the US Food and Drug Administration, with many others at various stages of the drug development pipeline. These drugs interfere with distinct steps in the viral replication cycle, resulting in a sharp drop of plasma viral RNA levels. However, single-drug therapy rapidly selects viral variants that carry mutations conferring resistance to the given drug, leading to a return of viral load to high levels. In contrast, modern combination therapy, which consists of the administration of at least three drugs from more than one class, can suppress viral replication for substantially longer times. This ongoing suppression is often accompanied by a substantial increase in CD4<sup>+</sup> counts. In this way, meeting the virological goal of anti-HIV therapy ideally entails meeting the immunological and clinical goals. Indeed, many clinical studies have shown a striking decrease in HIV-induced morbidity and mortality since the introduction of combination therapy (e.g. [Mocroft et al., 1998](#); [Hogg et al., 1998](#); [Palella et al., 1998](#)), to a degree that HIV-1 infection is now considered by many as a “treatable chronic disease” ([Simon et al., 2006](#)).

Despite these successes, even modern combination therapy can only delay the emergence of drug-resistant variants in most patients. Resistance will render the current regimen ineffective, leading again to high levels of viremia. When faced with such a therapy failure, the treating physician has to select a new – and preferably optimal – regimen. This task is highly complex due to increasing numbers of available antiretroviral drugs, significant cross-resistance and the likely presence of archived drug-resistant viral variants selected by previous regimens (discussed in detail in Chapter 2). Parameters with potential impact on treatment decisions which are typically obtained before choosing a new regimen include plasma viral load, CD4<sup>+</sup> cell count, viral genotype (determined by sequencing the relevant parts of the viral genome), phenotypic resistance, and prior treatment history. Other factors of relevance include tolerability, toxicity, the ability to preserve future treatment options, and data on drug metabolism. Currently, physicians interpret these complex and heterogeneous data using a subjective mix of information from guidelines, experience, and intuition. In contrast, model-based anti-HIV therapy aims to provide decision support which is justified by objective, reproducible, and data-driven criteria and studies. It is now being realized that “computational tools will be essential as exploratory and interpretation systems in order to obtain a better support of clinical decisions concerning both the prediction and the evolution of drug resistance” ([Carvajal-Rodriguez, 2007](#)).

Currently, HIV-1 infection entails the necessity of a life-long drug therapy. The main obstacle to eradicating HIV from a patient seems to lie in the viral ability to build a latent reservoir within resting memory CD4<sup>+</sup> cells. This means that both wild-type and drug-resistant viruses circulating for significant periods will get stored and can re-emerge if therapy is stopped or changed. Due to the lack of viral replication, virus stored in resting cells cannot be reached by current anti-HIV therapies. Several approaches to target these reservoirs are under investigation, but their clinical feasibility

---

remains to be determined.

### HIV: test for drug design methodologies in the past – test for drug administration methodologies in the future

Before 1996, only one class of drugs was available for anti-HIV therapy. The years 1995–1996, with the approval of the first protease inhibitors represented a triumph for anti-HIV therapy: it was found that the simultaneous administration of the novel drugs with reverse transcriptase inhibitors could prolong the development of resistance substantially. David Ho, one of the main proponents of the new combination therapy, was even voted *TIME* magazine's 1996 Man of the Year.

However, protease inhibitors did not only represent a triumph for antiviral therapy. They also represent one of the first successes of the then-new paradigm of structure-based (or: rational) drug design. Historically, new lead compounds had been discovered by chance or by mass screening of compound libraries. Not surprisingly, the success rate of these approaches was very low. In contrast, the first protease inhibitors – saquinavir, zidovudine, and zalcitabine (whose development started in 1989, with approval in late 1995 and early 1996) – had been derived by excessive use of all rational tools available at that time, including three-dimensional modeling studies and database searches for novel scaffolds (Böhm et al., 1996, pp. 494ff; in German). Due to these efforts, the case of the first protease inhibitors has been termed “a test for drug design methodologies” (West and Fairlie, 1995).

With the historical focus on rationalizing “drug discovery” into a process of “drug design”, considerably less attention has been devoted to modernizing the later stages of drug development, and their clinical use. It was only in 2004 that the United States Food and Drug Administration drew attention to the fact that “a new product development toolkit – containing powerful new scientific and technical methods such as [. . .] computer-based predictive models [. . .] is urgently needed to improve predictability and efficiency along the critical path from laboratory experiment to commercial product” (FDA, 2004). Among several other challenges, “new medical technologies, including [. . .] individualized drug therapies” (FDA, 2004) are highlighted prominently.

The goal of individualizing drug therapies can be paraphrased as “customized medical care for each patient's unique condition” (van der Greef et al., 2006). Ideally, such customized care would involve “a system of patient evaluation that would tell clinicians the correct drug, dose or intervention for any individual before the start of therapy” (Nicholson, 2006). In general, individualized medicine includes not only intervention, but also the evaluation of an individual's predisposition to diseases prior to its potential onset. Until recently, the pharmaceutical industry itself has not been a driving force behind individualizing therapies, largely due to uncertainty of how the traditional model of mass-market block buster drugs will be affected by the new treatment paradigm (Abrahams et al., 2005). However, with recent advances in metabolomics and pharma-

cogenomics, and with the \$1000 personal genome much faster in sight than initially anticipated (Service, 2006), it is obvious that future medicine will be strongly individualized. Implications of the paradigm shift away from the outdated “one size fits all” model of drug discovery, drug development, and healthcare infrastructure are discussed in (Jain, 2006; Meyer et al., 2002; Abrahams et al., 2005).

Most importantly, in individualized medicine, “combination drug therapies with individualized optimization are likely to become a major focus” (van der Greef et al., 2006). There is no doubt that among all diseases, the treatment of HIV is most advanced with respect to these challenges. Once again, similar as in the case of protease inhibitors mentioned above, challenges posed by HIV are leading the way into a new era, and serving as a test for methodologies. In other words, a challenge for drug design has now become a major challenge for their optimal use as well. While most individualization is performed with respect to viral, rather than host characteristics, extending the approaches to more classical host pharmacogenomics data, such as genetic variation in the cytochrome P450 enzyme family (affecting drug metabolism), or information about host Major Histocompatibility Complex haplotype, is methodologically straightforward, once enough data become available (e.g. Section 6.3 describes the inclusion of information about the allelic state of a certain cell-surface receptor relevant in HIV cell entry).

The challenge of selecting an appropriate drug combination from several available drugs based on pharmacogenomics and resistance data is now becoming increasingly common in many other viral, bacterial, fungal and parasitical diseases (Nosten and Brasseur, 2002) that take a high toll on our society. In this context, the experiences made with anti-HIV therapy will serve as “a model for the feasibility of treating chronic viral infections, [. . .] such as HBV and HCV” (Knipe et al., 2007, p.2198). Moreover, the treatment of cancer is also on the verge of individualized (combination) therapy (Ozols et al., 2007; Daly, 2007; Fonseca and Stewart, 2007). For example, in lung cancer, several genetic alterations in the *EGFR* gene have been found to be associated with how well patients will respond to a given drug (reviewed in Ozols et al., 2007).

## Outline

The goal of this thesis is to advance, on a quantitative basis, our understanding of HIV evolution in the presence and absence of drug, of the relation between genotype, phenotype, and response to therapy, and of the genetic basis of coreceptor usage. As should be clear from the introductory discussion, these questions are critically relevant as building blocks of rational, individualized anti-HIV therapy.

We begin in Chapter 2 with a brief review of the essential biomedical background on HIV-1 and AIDS, as a preparation for the main chapters of this thesis. Chapter 3 is devoted to an integrated discussion of scientific modeling and statistical learning. Supervised statistical learning methods will play a major role throughout this thesis.

In our discussion, they will appear as extremely empirical (i.e. non-explanatory) forms of scientific models. A section discussing the various modeling strategies taken in this thesis concludes the chapter.

Chapter 4 is devoted to the role of evolution. A classification system for fitness landscapes and its role in influencing the evolutionary dynamics of populations are studied using simulated data (Section 4.1). After these theoretical considerations, we shall turn our attention to evolutionary modeling based on data from clinical practice. To date, little is known about the dynamics of mutation disappearance in the absence of drug. Does it invert exactly the patterns observed during the emergence of resistance under drug pressure, or does it follow different principles? This question is addressed in Section 4.2. Finally, in Section 4.3, we show how to incorporate evolutionary modeling into approaches for predicting drug resistance or virological response from genotype.

Section 4.3 also serves as a bridge from Chapter 4 to Chapter 5, in which we focus on the various manifestations of resistance and their interpretation. We start with a comprehensive overview of the field of resistance interpretation algorithms (5.1). We then describe novel approaches for identifying resistance mutations and for characterizing their co-occurrence behavior, followed by an application to nucleoside and non-nucleoside reverse transcriptase inhibitors (5.2). The question of whether clinical resistance should be predicted by direct correlation of genotypes with virological response, or by using predicted phenotypes as an “intermediate” step has been subject to some controversy. We study this question in a large-scale evaluation in Section 5.3. Concluding the chapter, we investigate potential benefits of incorporating an additional phenotypic property of the virus – “fitness”, or, more accurately, replication capacity – into models for virological response prediction (5.4).

Finally, in Chapter 6, we focus on the process of viral cell entry. The entry process is initiated by attachment of the viral envelope protein to the CD4 receptor. However, fusion of viral and cell membranes depends on the additional interaction with a coreceptor (mainly CCR5 or CXCR4). The coreceptor choice of a viral population is not only tightly linked to disease progression, but also relevant for the treatment with a novel class of anti-HIV drugs called coreceptor antagonists, as reviewed in Section 6.1. We describe a comprehensive comparison of different methods for predicting coreceptor usage from (part of) the sequence of the viral envelope protein (Section 6.2). It will be apparent that the reliability of purely sequence-based prediction is not satisfactory on clinical sequence data obtained from population-based sequencing. For this reason, we propose to combine sequence-based models with clinical and host markers, and data on the heterogeneity of the viral population *in vivo*, to improve the prediction of coreceptor usage in a clinical setting (6.3). Finally, we describe an approach to improving predictive reliability by structural modeling (6.4).



---

# Anti-HIV Therapy

In this chapter, we provide the necessary background on HIV/AIDS. Many important aspects not immediately relevant for this thesis had to be omitted. Excellent monographs on virology (Knipe et al., 2007; Flint et al., 2000), anti-HIV therapy (Hoffmann et al., 2006; Butera, 2005), drug resistance (Clavel et al., 2004; Geretti, 2006), and modeling for anti-HIV therapy (Crandall, 1999; Nowak and May, 2000; Rodrigo and Learn, 2001; Tan and Wu, 2005) are available.

## 2.1 Epidemiology

HIV-1 belongs to the genus *Lentivirus*, as part of the family of retroviruses (*Retroviridae*). Two rather different forms of HIV have been described, HIV-1 and HIV-2, the former being the more common and more pathogenic, and the subject of this thesis. Based on genetic similarity, HIV-1 viruses are classified into three major groups: M (the “main”, or pandemic group), O, and N (Figure 2.1), which have been implicated with separate cross-species transmission events in the early 20<sup>th</sup> century (Hahn et al., 2000; Korber et al., 2000; Sharp et al., 2001). The chimpanzee communities from which groups M and N originate were recently discovered in two geographically distinct regions in southern Cameroon (Keele et al., 2006). As shown schematically in Figure 2.1, group M consists of several distinct clusters, called subtypes, and many circulating and unique recombinant forms, which carry genetic material from two or more subtypes. While subtype C accounts for approximately 55% of all HIV-1 infections worldwide, subtype B dominates in Europe, America, and Australia.

In the context of this thesis, the genetic diversity of HIV-1 is only relevant regarding potential implications on computational procedures. While the impact of subtype diversity on vaccine design is obvious, the impact on response to anti-HIV therapy, resistance evolution and disease progression remains controversial (reviewed in Camacho, 2006). However, and most importantly, it is very plausible that models derived from databases (mostly consisting of subtype B sequences) by statistical learning are

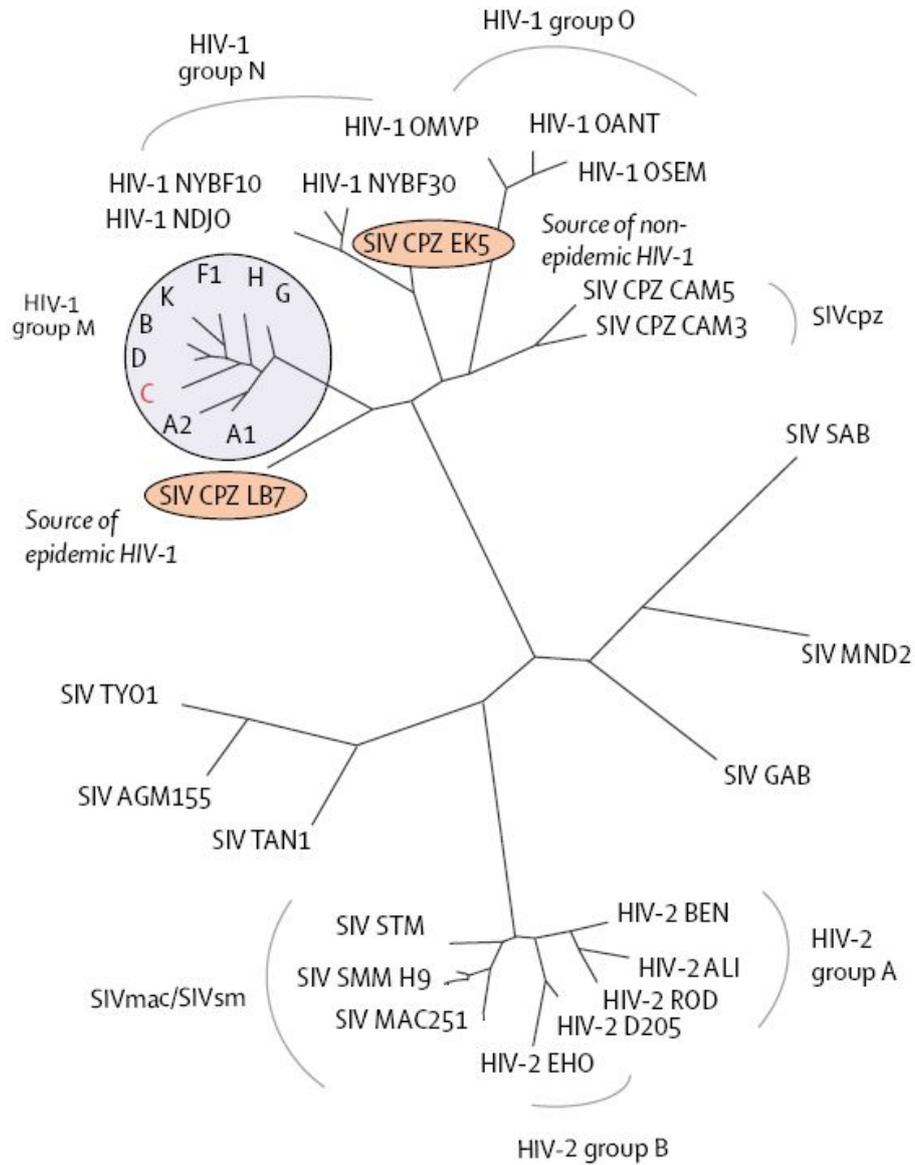


Figure 2.1: Phylogenetic relations among HIV and SIV groups and subtypes. From (Simon et al., 2006).

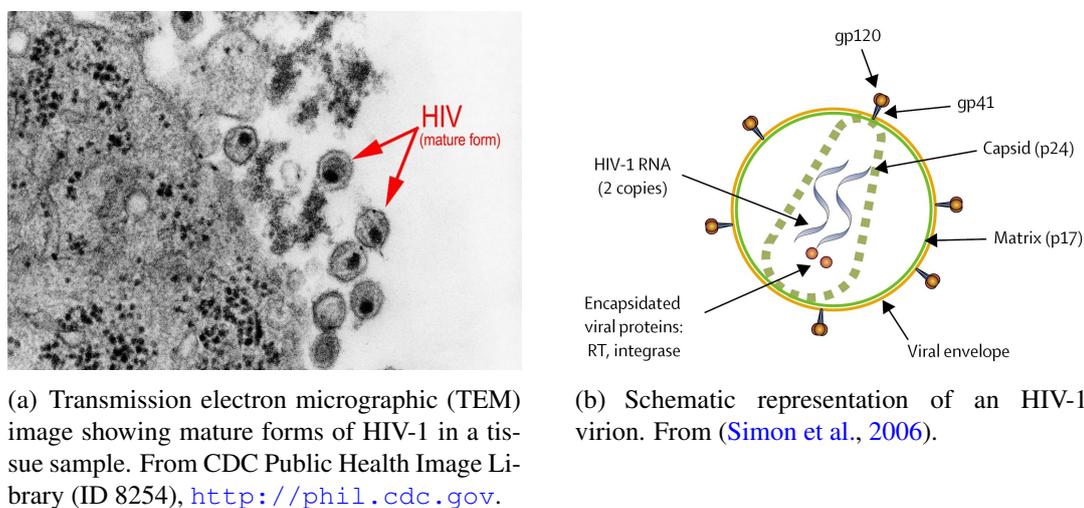


Figure 2.2: Structure of the HIV virion.

less reliable for other subtypes. In fact, for this very reason, it has been suggested that “results using automated interpretation systems should be analysed and corrected by an expert virologist before a final report is sent to the clinician” (Camacho, 2006). For example, in predicting coreceptor usage from the sequence of the viral envelope protein (cf. Chapter 6), it has been reported that methods trained on subtype B sequences perform poorly on subtype C data. As a consequence, a subtype-specific prediction approach has been proposed (Jensen et al., 2006).

## 2.2 Structure of virion, genome, and viral proteins

HIV-1 virions are spherical particles with a diameter of about 100 nm (Figure 2.2(a)). A schematic representation is shown in Figure 2.2(b). The virion is surrounded by an envelope, consisting of a lipid bilayer. The prominent surface spikes are heterotrimeric complexes consisting of three copies of the viral envelope protein gp120 (cf. below for details on the viral proteins), along with three copies of the viral transmembrane protein gp41. Another spherical layer, consisting of copies of the p17 (matrix) protein, is located immediately below the envelope membrane. The matrix layer serves to stabilize the viral envelope. The actual core (or capsid) of the virion is located beneath the matrix layer. The capsid is composed of about 2,000 copies of the viral p24 protein.

The viral genomic information is sheltered within the capsid. It consists of two single-stranded copies of RNA, each about 9.7 kb in length. The two strands are identical or almost identical. Along with the genetic material, several viral proteins are packaged within the core. Two of them are of particular relevance in the context of this thesis. Reverse transcriptase (RT) is the key enzyme for catalyzing the conversion

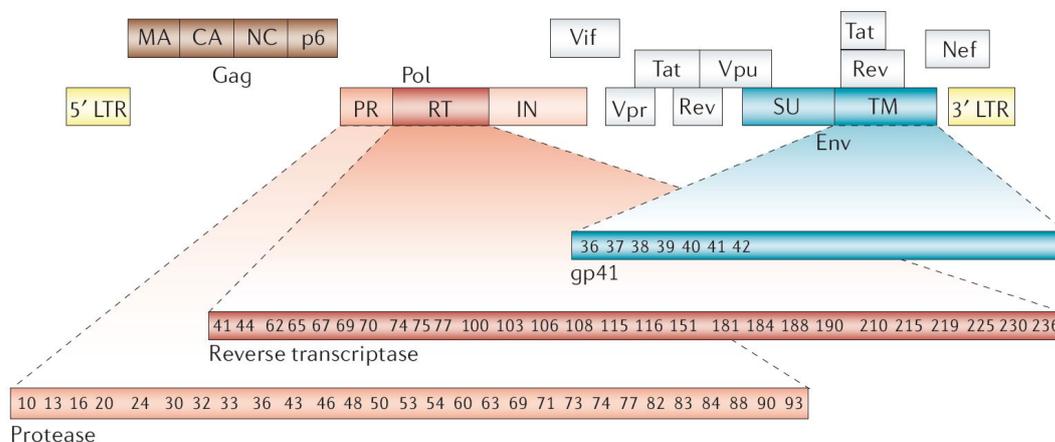
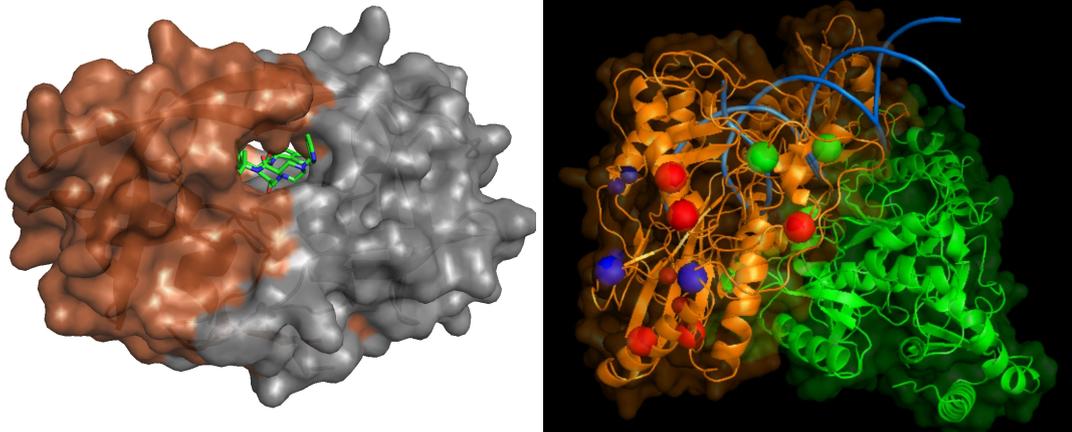


Figure 2.3: The HIV-1 genome. Currently approved antiretroviral drugs are targeted against the viral protease, reverse transcriptase, and the transmembrane protein gp41. Resistance mutations in the genetic regions encoding these proteins are highlighted. From (Lengauer and Sing, 2006).

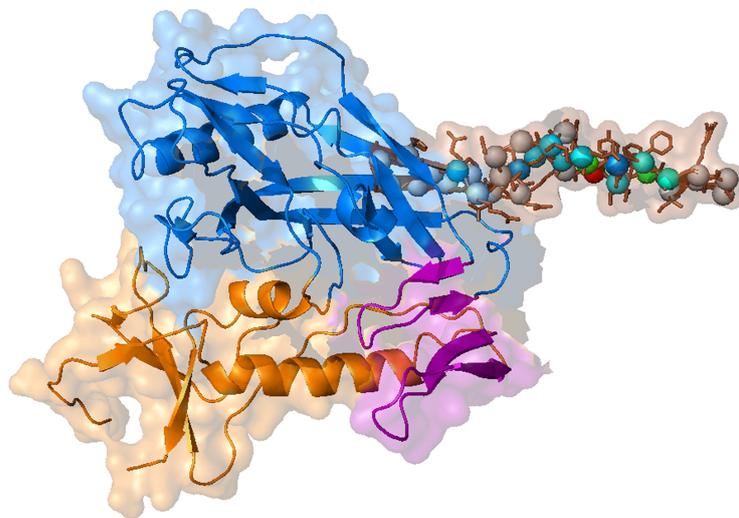
of viral RNA back to DNA. Without this RT-mediated conversion, the viral genome could not be “rewritten” in a form appropriate for integration into the human genome. The viral protease (PR) cleaves viral polyproteins into functional proteins. Without the cleavage activity of PR, newly produced viral particles would remain non-infectious.

The viral genome contains nine genes that code for 15 proteins (Figure 2.3). The genes occupy all three reading frames, with considerable overlap. The overall genomic organization  $5' - gag - pol - env - 3'$  is common to all retroviruses. In this thesis, we are exclusively concerned with the proteins PR, RT, and gp120. These are encoded by the *pol* (PR, RT) and *env* (gp120) genes, respectively. In the remainder of this section, we shall summarize the main characteristics of the three enzymes.

HIV protease (Figure 2.4(a)) is a homodimeric enzyme consisting of two identical symmetrical subunits of 99 residues each. A cavity is located in the middle of the enzyme, forming the substrate-binding cleft. The active site of the enzyme lies at the bottom of this cleft. From the top, the cavity is covered by two mobile flaps which allow the substrate to enter or leave. In HIV-1, and in several other viruses (Flint et al., 2000, p.467–471), newly produced virions budding from an infected cell are non-infectious. This is because HIV produces most of its proteins as polyproteins which have to be cleaved into functional proteins before virus particles can mature to become infectious. The polyproteins cleaved by HIV protease are Gag (encoding structural proteins of the virus) and Gag-Pol (encoding the protease itself, the reverse transcriptase, and the integrase). Protease-mediated cleavage of these precursors proceeds at different cleavage sites in a well-defined order. As mentioned in Chapter 1, protease inhibitors (cf. also Section 2.4) did not only manifest a break-through in an-



(a) HIV protease (with the two subunits colored differently) in complex with the inhibitor *Indinavir* (PDB ID 1sdt).  
(b) Reverse transcriptase (with p51 and p66 colored differently) in complex with DNA (PDB ID 1rdt). Novel mutations reported in Section 5.2 are highlighted.



(c) The envelope protein gp120. The “outer” and “inner domains” are shown in blue and orange, respectively. The bridging sheet is colored in purple. The V3 loop is shown in grey. V3 residues associated with increased propensity for X4 are colored (cf. Section 6.2).

Figure 2.4: The viral proteins of relevance for this thesis. Rendered with the Pymol software.

tiviral therapy, but also represented one of the first successes of rational drug design.

Reverse transcription is the “hallmark of the retroid viruses” (Coffin et al., 1997). The proposal that genetic information could be copied into DNA, and that this DNA could become integrated into the genetic material of cells was originally considered “heresy”<sup>1</sup> by most scientists because it violated the widely accepted “central dogma of molecular biology” (Crick, 1970). The dogma states that information flow in nature always occurs from DNA to RNA, and never in the opposite direction. However, in the late 1960s, an enzyme – appropriately named *reverse transcriptase* – was discovered in certain tumour viruses that could indeed produce DNA copies from RNA (Temin and Mizutani, 1970; Baltimore, 1970). In 1975, the Nobel Prize in physiology or medicine was awarded to Howard Temin and David Baltimore for these discoveries. The reverse transcriptase of HIV (Figure 2.4(b)) is a heterodimer consisting of two subunits named p66 and p51. The p66 subunit contains two domains, polymerase (N-terminal, 440 residues) and RNase H (C-Terminal, 120 residues). The p51 subunit consists of the polymerase domain from the p66 subunit. However, the spatial arrangement of the two analogous domains is very different, and they are also associated with different functions in the enzyme. The molecular process of reverse transcription is very complex and not relevant in the context of this thesis (Flint et al., 2000, p.200-214). In our context, the single most important fact is that unlike other DNA polymerases, the RT lacks a proofreading mechanism. This results in a high number of errors during reverse transcription, and constitutes the source for the extremely high variability of HIV *in vivo*, which presents a major obstacle to successful therapy or vaccination (cf. Section 2.5).

The envelope protein gp120 is the building block (along with the transmembrane protein gp41) for the surface spikes protruding from the envelope membrane. It consists of about 480 residues, although its length is very variable compared to protease or RT, where insertions or deletions occur much less frequently. The enzyme is composed of three domains (Figure 2.4(c)). The “inner” domain (shown in orange) covers the N- and C-terminal regions of gp120. It consists of two  $\alpha$  helices, two  $\beta$  strands and a small  $\beta$  sandwich next to the terminals<sup>2</sup>. The “outer” domain (blue) is a stacked double barrel structure covering the middle of the sequence. The inner and outer domains are connected by a “bridging sheet” (purple) which consists of four antiparallel strands. In Chapter 6, we focus on the grey structure extending from the outer domain as a loop. It is encoded by a highly variable region of the *env* gene of length 33 to 37. The region usually starts with the amino acid motif CTR or CIR, ends with AHC or AYC, and contains a conserved GPG motif in the middle (mapping to the tip of loop in the structure). Starting from the N-terminus of the gp120 primary structure, this

---

<sup>1</sup>Press release for the Nobel Prize in physiology or medicine 1975. From [http://nobelprize.org/nobel\\_prizes/medicine/laureates/1975/press.html](http://nobelprize.org/nobel_prizes/medicine/laureates/1975/press.html).

<sup>2</sup>Frequent structural motifs in proteins are reviewed in (Branden and Tooze, 1999, p. 13–88).

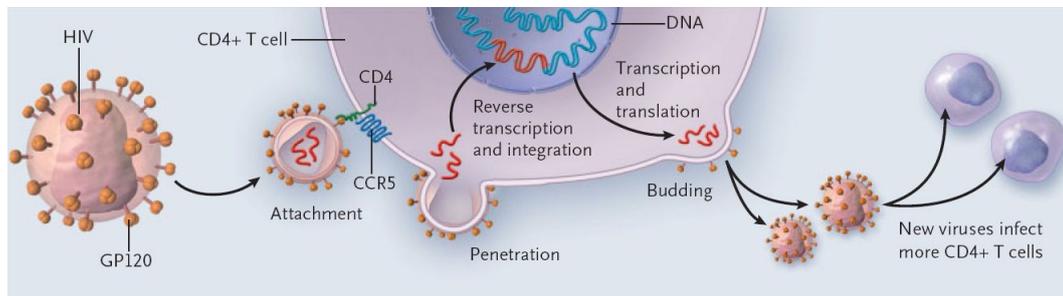


Figure 2.5: The mechanism of HIV replication. From (Markel, 2005).

region is the third highly variable region (although the “constant” regions are far from constant). For these reasons, it has been called the third hypervariable (V3) region, and the structural motif the V3 loop. As discussed in detail in Section 6.1, the V3 loop has been implicated as one of the main determinants of differential viral coreceptor usage.

## 2.3 Replication and pathogenesis

HIV-1 replication is a complex process whose outcome and duration depend on a number of factors, including target cell type and cell activation. As shown in Figure 2.5, HIV cell entry is mediated by consecutive interaction with the CD4 cell surface receptor and one of the two major coreceptors CCR5 and CXCR4. After binding to CD4, a conformational switch in the surface protein gp120 of HIV reveals the coreceptor binding site, most notably, the third hypervariable loop region V3. The V3 loop is considered to be a major viral determinant for coreceptor specificity, and will be a focus of Chapter 6. After successful binding to the coreceptor, fusion of the viral and host cell membranes takes place (Berger et al., 1999). The coreceptor selectivity of the viral population is of central pathological and clinical importance. In newly infected patients, CCR5-using (R5) variants dominate. However, in around 50% of the patients, CXCR4-using (X4) variants appear during later stages of the disease characterized by progression towards AIDS (cf. Chapter 6).

During cell entry, the virion’s envelope and matrix layer are lost. However, the capsid stays intact until the genome has been reverse transcribed to DNA. In the meantime, the capsid moves to the nucleus of the cell. There it opens and the viral DNA (called provirus) enters the nucleus, along with other viral proteins. With the help of integrase, another viral protein, the provirus is inserted into the human genome. This marks the “turning point” (Simon et al., 2006) from which a cell is irreversibly transformed into a potential virus producer. Viral mRNA is then produced by the normal transcription machinery of the cell. The mRNAs are transported out of the nucleus and translated into amino acid sequences consisting of multiple uncleaved proteins.

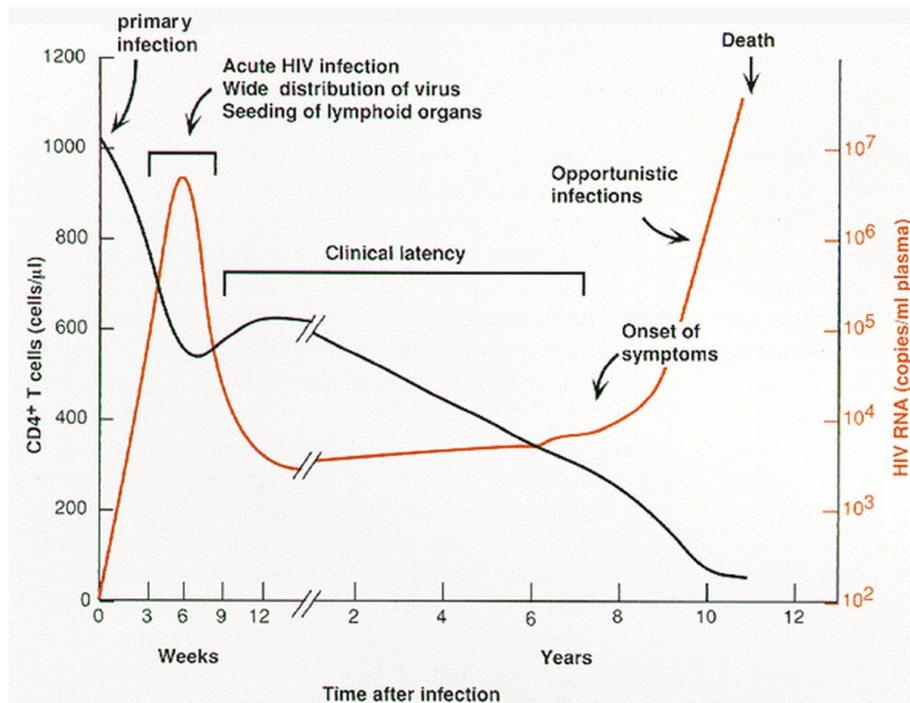


Figure 2.6: Schematic course of viral and  $CD4^+$  cell dynamics during HIV infection. From (Coffin et al., 1997)

These polyproteins are then shipped to the cell membrane where they are packed into new virus particles. Finally, the newly assembled virions bud from the cell, taking part of the cell membrane as their envelope. Meanwhile, inside the virion, the protease cleaves the non-functional polyproteins into functional proteins. Each of these steps is the result of complex, temporally and spatially highly coordinated interactions, whose details are still only partially understood and beyond the scope of this thesis.

As mentioned in Chapter 1, HIV-1 pathogenesis is characterized by a gradual destruction of the naive and memory  $CD4^+$  T-lymphocyte populations, with AIDS as the end stage (reviewed in Knipe et al., 2007, p.2187–2214). Figure 2.6 shows a schematic representation of  $CD4^+$  cell and virus dynamics in peripheral blood during the course of infection. According to this picture, HIV infection is commonly divided into three phases. The *acute phase* of infection is characterized by high plasma viral load and low  $CD4^+$  cell counts. During peak levels of viral load, clinical symptoms can be manifest. After several weeks, the acute phase levels off into a *latent (asymptomatic)* phase, characterized by  $CD4^+$  cell counts and viral load at relatively stable levels and the absence of clinical symptoms. However, these stable levels are the result of a steady state of virus and  $CD4^+$  cell dynamics in which rapid, continuous turnover takes place

within these populations (reviewed in Douek et al., 2003).<sup>3</sup> The latent phase can last between two and 20 years with slowly decreasing CD4<sup>+</sup> cell counts, until eventually the immune system collapses, leading to fatal immunodeficiency, sharp increases in viral load, and the emergence of AIDS-associated morbidities in the *clinical phase* of infection.

The three-stage description of HIV pathogenesis suggests immune depletion to be a direct consequence of HIV replication in the CD4<sup>+</sup> cell population. However, the extensive amount of viral replication throughout the course of infection clashes with the rather slow deterioration of CD4<sup>+</sup> cells (reviewed in Simon and Ho, 2003). In fact, HIV RNA levels have been found to predict the rate of CD4<sup>+</sup> cell decline only minimally, thus “challenging the concept that the magnitude of viral replication (as reflected by plasma levels) is the main determinant of the speed of CD4<sup>+</sup> cell loss” (Rodriguez et al., 2006). The inability to resolve this central question of HIV pathogenesis has stimulated heated debate within the scientific community, mostly concerning the relative role of direct and indirect effects of viral replication (reviewed in Douek et al., 2003). Recently, substantial attention has been devoted to the fact that most data originate from peripheral blood. Such data can be easily obtained, but may provide only a limited view of how HIV affects the immune system as a whole. In fact, a rapid (within days) and dramatic depletion of both active and memory CD4<sup>+</sup> T cells in gut-associated lymphoid tissues (which contain 70% of a body’s lymphocyte population) has been observed (Guadalupe et al., 2006). This depletion seems to remain mostly irreversible (Mehandru et al., 2006), while the CD4<sup>+</sup> cell population in the peripheral blood often returns to normal levels in response to antiretroviral therapy. These data have been interpreted as establishing “that it is the virus in the acute phase of the disease rather than immune activation in the chronic phase that is responsible for the bulk of CD4<sup>+</sup> T cell depletion” (Brenchley et al., 2006a) and that the major battlefield might be located in gut-associated lymphoid tissue rather than in peripheral blood. However, this conclusion does not explain why sooty mangabeys infected with SIV show a similar gut depletion, but almost never progress to immunodeficiency (reviewed in Grossman et al., 2006). More recent work provides evidence that the compromised lymphoid tissue in the gut may lead to circulating microbial products in HIV but not in SIV infection, thus suggesting a possible explanation for differences in human versus simian hosts. This bacterial translocation has been suggested as a major cause of HIV-related systemic immune activation (Brenchley et al., 2006b). While important details remain poorly understood, the recent work could lead to a “massive reappraisal” (Check, 2007) of the mechanisms of HIV pathogenesis.

---

<sup>3</sup>Sompayrac (2002, p.69) uses more pictorial language in his description of this disease stage in which “huge amounts of virus are produced, large numbers of CD4<sup>+</sup> T cells are continuously killed and replaced, and in which the host immune system is engaged in a heroic battle to control and eliminate the virus.”

## 2.4 Anti-HIV drugs

As mentioned in Chapter 1, antiretroviral therapy is based on the hypothesis that maximal suppression of viral replication will prevent or delay immune deterioration and avoid HIV-associated morbidity and mortality. Suppression is achieved by interfering with the viral replication cycle at various steps. As of April 2007, 21 compounds from different mechanistic classes have been approved by the US Food and Drug Administration for anti-HIV therapy and two further compounds from entirely new classes are expected to enter the market before the end of the year (Opar, 2007). In this section, we shall give an overview of the different drug classes and their mechanism of action.

Reverse transcriptase inhibitors comprise more than half of the currently approved anti-HIV drugs (Table 2.1). They interfere with the process of reverse transcription via two different mechanisms. *Nucleoside reverse transcriptase inhibitors* (NRTIs) are analogues of deoxynucleotides, the natural building blocks of DNA. However, unlike deoxynucleotides, NRTIs lack a 3'-OH group which is needed for primer elongation. Thus, once incorporated into a growing cDNA strand, they will disrupt reverse transcription by acting as chain terminators. NRTIs have formed the backbone of antiretroviral therapy since the introduction of zidovudine as the first anti-HIV drug in 1987 (Table 2.1).

*Non-nucleoside reverse transcriptase inhibitors* (NNRTIs) act as non-competitive inhibitors of reverse transcriptase by occupying a hydrophobic pocket close to, but distinct from the catalytic domain of RT. The binding of the inhibitor induces a structural alteration in RT, thus blocking its polymerase activity. In face of their high potency coupled with low toxicity, NNRTIs are preferred agents for first-line regimens and for preventing mother-to-child transmission of HIV.

*Protease inhibitors* (PIs) mimic the natural substrates of the viral protease. They occupy the active site of the enzyme, thereby preventing HIV protease from processing the Gag and Gag-Pol precursor polyproteins (cf. Section 2.2) into functional proteins. Unlike RT inhibitors, PIs do not prevent virus from persistently infecting cells by interrupting the replication cycle before the irreversible “turning point” (cf. Section 2.3) of integration. Rather, they prevent infected cells from producing infectious particles by preventing maturation of budding virions. Due to their relatively poor bioavailability, most PIs are co-administered with a low dose of ritonavir, a protease inhibitor that inhibits cytochrome  $P_{450}$ -mediated metabolism of these drugs. PIs have proven highly effective in both initial and subsequent treatment regimens.

During the last few years, the novel drug class of *entry inhibitors* has rapidly moved from laboratory to clinical evaluation (reviewed in Derdeyn and Hunter, 2005). Blocking extra- rather than intracellular interactions, entry inhibitors need not be transported across the cell membrane. Different classes of entry inhibitors targeting distinct steps in the process of HIV entry (cf. Section 2.3), including attachment, coreceptor binding,

Table 2.1: Anti-HIV drugs with year of FDA approval. The coreceptor antagonist maraviroc and the integrase inhibitor raltegravir are expected to receive FDA approval in 2007.

Generic name	Abbreviation	Trade name	FDA approval
Nucleoside reverse transcriptase inhibitors (NRTIs)			
zidovudine	ZDV, AZT	Retrovir	1987
didanosine	ddI	Videx	1991
zalcitabine	ddC	Hivid	1992
stavudine	d4T	Zerit	1994
lamivudine	3TC	Epivir	1995
abacavir	ABC	Ziagen	1998
tenofovir	TDF	Viread	2001
emtricitabine	FTC	Emtriva	2003
Non-nucleoside reverse transcriptase inhibitors (NNRTIs)			
nevirapine	NVP	Viramune	1996
delavirdine	DLV	Rescriptor	1997
efavirenz	EFV	Sustiva	1998
Protease inhibitors (PIs)			
saquinavir	SQV	Fortovase, Invirase (SQV+RTV)	1995
ritonavir	RTV	Norvir	1996
indinavir	IDV	Crixivan	1996
nelfinavir	NFV	Viracept	1997
fos-/amprenavir	FPV/APV	Lexiva/Agenerase	2003/1999
lopinavir	LPV	Kaletra (LPV+RTV)	2000
atazanavir	ATV	Reyataz	2003
tipranavir	TPV	Aptivus	2005
darunavir	TMC114	Prezista	2006
Fusion inhibitors (FIs)			
enfuvirtide	ENF, T-20	Fuzeon	2003
Coreceptor antagonists			
maraviroc	(UK-427,857)	(NA)	(2007)
Integrase inhibitors			
raltegravir	(MK-0518)	(NA)	(2007)

and membrane fusion, have arrived at various stages of clinical development. In 2003, the *fusion inhibitor* enfuvirtide has become the first entry inhibitor to be approved by the FDA (reviewed in [Matthews et al., 2004](#)). The work described in Chapter 6 is motivated by the expected approval (reviewed in [Opar, 2007](#)) of the first member of a novel subclass of entry inhibitors called *coreceptor antagonists* ([Westby and van der Ryst, 2005](#)). These drugs prevent HIV entry by blocking one of the cell-surface receptors used by HIV as a coreceptor. Unlike previous anti-HIV drugs, coreceptor antagonists target host, rather than viral proteins.

Many other drugs are under investigation. Another drug class whose first member is expected to become approved in 2007 are integrase inhibitors. With two new classes of anti-HIV drugs poised to enter the market in 2007, “a watershed in HIV treatment, second only to the introduction of cocktail therapy more than one decade ago” ([Opar, 2007](#)) has been anticipated. Currently, about 20 anti-HIV drugs are in various stages of clinical development, most of them belonging to entirely new classes of drugs ([Opar, 2007](#)), and many more are in pre-clinical stages of the drug development pipeline. Moreover, novel drugs (e.g. tipranavir or darunavir) from “traditional” classes (NRTI, PI, NNRTI) are being designed so as to minimize cross-resistance (cf. Section 2.5) with older drugs from the same class.

Besides antiretroviral drug therapy – the focus of this thesis – several other anti-HIV strategies are investigated, including immune-based therapies, RNA interference, or the clearance of latent reservoirs (all reviewed in [Butera, 2005](#)). While fully suppressive antiviral therapy in combination with effective clearance of latent reservoirs could provide a cure for HIV, the complete eradication of the disease would crucially depend on the development of a preventive vaccine. Indeed, the availability of such a vaccine was proclaimed within two years after the discovery of HIV in 1983/84 ([Smith, 2003](#)). The eradication of smallpox through vaccination programs which had been declared by the World Health Organization in 1979 served as an encouraging example at that time. However, over twenty years later, and with more money spent than for any other vaccine effort in history ([Cohen, 2005](#)), the race towards an HIV vaccine has proved nothing but a “Sisyphean onslaught of disappointments” ([Markel, 2005](#)). Still, the potential of thwarting millions of new HIV infections each year more than justifies the ongoing high investments in HIV vaccine research. In fact, the 125-year celebration issue of the journal *Science* has selected the question “Is an effective HIV vaccine feasible?” ([Cohen, 2005](#)) as one of 25 scientific questions of outstanding impact until 2030 ([Kennedy and Norman, 2005](#)).

## 2.5 Evolution, drug resistance, combination therapy

Drug resistance refers to the ability of a viral population to replicate in the presence of drug. The first accounts of HIV drug resistance appeared shortly after approval of

the NRTI zidovudine as the first anti-HIV compound (Larder et al., 1989). In the same year, mutations in reverse transcriptase were identified as a cause for reduced susceptibility (Larder and Kemp, 1989), establishing a link between viral genotype and clinical resistance. Since then, the mechanisms of HIV drug resistance have been studied extensively. In general, resistance mutations occur in regions of the viral genome coding for the drug target.<sup>4</sup> They alter the structure and/or chemical properties of the targeted enzyme such that its molecular function is no longer inhibited. Two different mechanisms of NRTI resistance are known (reviewed in Marcelin, 2006): the improved differentiation between real nucleosides and nucleoside analogues, and the excision of already incorporated analogues in a process called “primer unblocking”. Resistance to NNRTIs involves mainly mutations in the RT pocket targeted by this class of drugs, reducing the affinity of the inhibitors for the enzyme. PI resistance is believed to be a more gradual process, with initial, or “primary” mutations appearing near the substrate-binding cleft of the enzyme. The structural alterations in the cleft lead to reduced susceptibility. During ongoing treatment, “secondary” mutations arise in other parts of the protease and can lead to drastic increases in resistance. Resistance to coreceptor antagonists can be manifest by a shift in coreceptor usage of the viral population, or by increased affinity to the current coreceptor (reviewed in Poveda and Soriano, 2006). Resistance to other drug classes is beyond the scope of this thesis. Due to the similarity of compounds within a given drug class, it is often observed that resistance selected by one compound also confers a certain degree of resistance to other, non-administered, compounds (“cross-resistance”).

Resistance to chemical intervention was nothing new in 1989. When the first reports on HIV drug resistance appeared, resistance had already been described in more than 500 species (reviewed in Palumbi, 2001). The time to emergence of resistance depends heavily on a species’ genetic variability and population turnover. Herbicide and insecticide resistance often evolves within time scales of about a decade from deployment (Palumbi, 2001). Resistance to antibiotics can evolve even faster, due to the increased variability and turnover of bacterial populations. For example, the first reports on penicillin resistance appeared within three years after its first use. Even faster bacterial evolution has been observed under controlled laboratory conditions (Lenski et al., 2003; Papadopoulos et al., 1999).

The highest rates of evolution, however, are observed within virus populations, at the border between the living and non-living world. This is especially true for RNA viruses, whose RNA polymerases (or, in the case of retroviruses, reverse transcriptases) typically lack proofreading-repair activity (in the form of a 3’ to 5’ exonuclease do-

---

<sup>4</sup>Resistance mutations can also appear in substrates, leading to enzyme-substrate co-evolution. For example, in the context of PI resistance, mutations in the protease-cleavage sites in Gag have been shown to emerge in response to therapy. Such mutations are considered as “compensatory” in the sense that they restore the catalytic activity of resistance enzymes, which is often reduced compared to wild-type enzymes.

main). For many RNA viruses, mutation rates have been estimated to lie in the range of  $10^{-3}$  to  $10^{-5}$  misincorporated nucleotides per nucleotides copied (reviewed in [Knipe et al., 2007](#), p.395). For HIV-1 with wild-type reverse transcriptase, a mutation rate of  $3.4 \times 10^{-5}$  has been estimated ([Mansky and Temin, 1995](#)). Given HIV's genome length of about 10 kbp, this means that an average of 0.1 to 1 misincorporations per template copied are to be expected. Owing to this powerful source of variability, RNA virus populations consist of complex distributions of mutant genomes – often termed quasispecies ([Domingo, 2005](#); [Domingo et al., 2001](#)) – rather than defined genomic sequences. Another source of variability is provided by recombination. Recombination in HIV is the formation of new genomes by template switching of the reverse transcriptase between two different genomes within a multiply infected cell. It has been reported to occur between two and three times per template copied ([Jetzt et al., 2000](#); [Zhuang et al., 2002](#); ?) — an even higher rate than the rate of nucleotide misincorporation. While nucleotide misincorporation and recombination provide the source of variability, the speed of adaptation depends heavily on the turnover of a population. In HIV, the turnover is extremely high, with a total daily virion production on the order of  $10^9$ , and a replication time of only one to three days (reviewed in [Simon and Ho, 2003](#)).

In practice, the extraordinary variability and turnover of HIV populations render monotherapy effectively useless: Clinical resistance to monotherapy has been observed as early as one or two weeks after infection ([Richman et al., 1994](#); [Condra et al., 1996](#)). In fact, in the case of NNRTIs, resistance has been reported to appear after only a single dose of monotherapy, which was given with the intention of preventing mother-to-child transmission ([Jackson et al., 2000](#); [Cunningham et al., 2002](#)). Moreover, follow-ups of early clinical trials revealed that the observed short-term benefits of monotherapy as compared to placebo therapy vanished over longer observation periods, with respect to either virological, immunological or survival endpoints (reviewed in [Butera, 2005](#)). In the light of these problems it was hypothesized from early on that “combination therapies that target different viral replicative sites likely will [...] help prevent drug resistance” ([Hirsch, 1990](#)). This hypothesis could be tested in the early 1990s, when two additional NRTIs were approved, didanosine and zalcitabine ([Table 2.1](#)). Several clinical trials (reviewed in [Butera, 2005](#), p.5–8) compared zidovudine monotherapy with the first combination therapies consisting of zidovudine and either didanosine or zalcitabine. The results of these studies were convincing enough to establish a new standard of care for anti-HIV therapy lasting until the mid-1990s: combination therapy with two NRTIs.

The real breakthrough for combination therapy occurred in the years 1995/1996, when the first drugs from the novel classes of PIs and NNRTIs became available. In several trials, combination regimens consisting of NRTIs plus one PI or NNRTI substantially outperformed purely NRTI-based combinations (reviewed in [Butera, 2005](#), pp. 9–16). These findings can be readily understood when considering the extraordi-

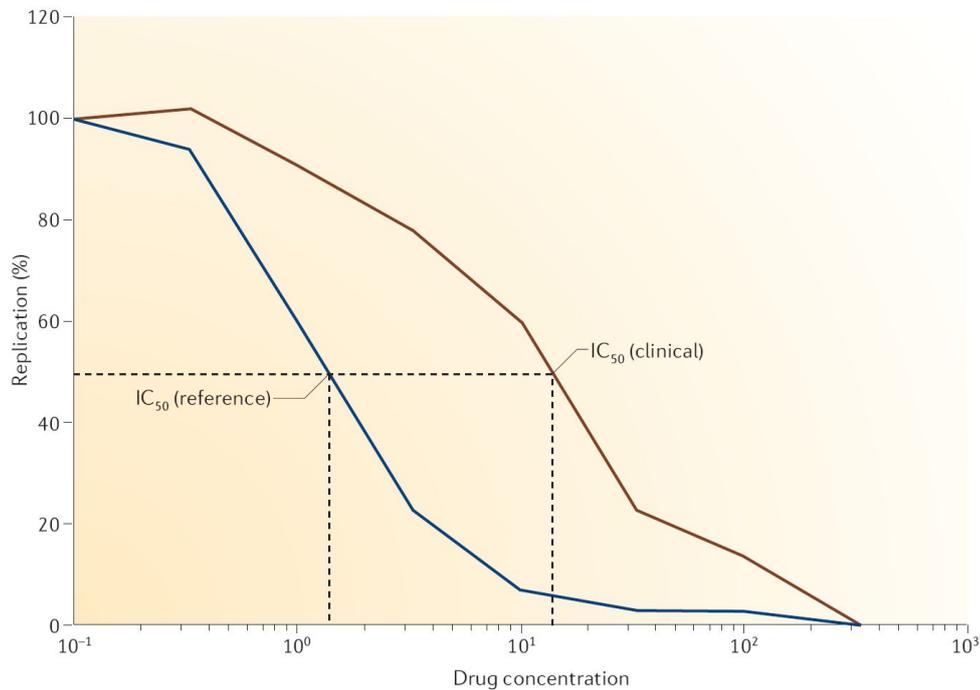


Figure 2.7: Dose-response curve for zidovudine, determined using an *in vitro* recombinant assay. The curve for the patient virus is shown in brown, whereas the reference virus is shown in blue. The quotient between the 50% inhibitory concentration ( $IC_{50}$ ) of the patient virus ( $14 \mu\text{mol}$ ) versus the reference virus ( $1.4 \mu\text{mol}$ ) represents a ten-fold change in susceptibility. Other parameters of the curve, such as the slope or the  $IC_{90}$  (the drug concentration that induces the reduction of viral activity to 10% of the activity in the absence of the drug), might have clinical relevance as well, but are not used routinely. From (Lengauer and Sing, 2006).

nary variability and turnover of HIV *in vivo* mentioned above: since variability is so high that virtually all possible point mutations along the HIV genome are produced on a day-by-day basis (Coffin, 1995), a certain level of resistance to any single drug is likely to be present in the population at any time. However, it is highly unlikely that a number of mutations conferring resistance to multiple classes of drugs are present on a single genome *a priori*. As a consequence, combination therapy with three or more drugs from two different classes quickly became the standard of care. Highly active antiretroviral therapy (HAART), as the strategy was called for almost a decade<sup>5</sup>, led to a dramatic reduction in HIV-related morbidity and mortality from 1995 on (e.g. Mocroft et al., 1998; Hogg et al., 1998; Palella et al., 1998). Moreover, the emergence of resistant variants can be delayed substantially with this strategy. Still, residual repli-

<sup>5</sup>Recently, the more neutral term “combination antiretroviral therapy” (cART) has been preferred.

cation will eventually lead to the production of multi-drug resistant virus and subsequent therapy failure. Combined with the existence of viral reservoirs (cf. Section 2.3), this means that HIV-infected individuals are in need of life-long combination therapy under changing regimens. Apart from the serious consequences for patients, life-long therapy also imposes a tremendous burden on public health: Based on an estimated life expectancy of 24.2 years from time to entering HIV care, the lifetime cost for treatment of a single HIV patient in the US has been estimated to be \$618,900, with 73% of this sum accounted for by the cost of antiretroviral drugs (Schackman et al., 2006).

As mentioned above, reduced susceptibility to drugs *in vivo* can be quantified by viral load measurements from blood samples. Viral load tests are usually given in units of “viral RNA copies per milliliter (cp/ml)” of blood plasma. For example, a value of  $10^5$  cp/ml indicates that 50,000 virions per ml are present (since each virion carries two copies of RNA). However, the link between mutations in the viral genotype and viral load is confounded by many additional parameters, such as inter-individual differences in immunology, drug metabolism, other host factors, or viral fitness. Moreover, given the ubiquitous use of combination therapy, it is far from straightforward to infer the contributions of individual drugs from such measurements. Thus, absolute levels of viral load are problematic as direct measures of viral resistance to a given drug. Rather, drug resistance should be quantified in isolation (i.e. in the absence of the confounding factors mentioned above) under well-controlled laboratory conditions. In phenotypic resistance assays (Schutten, 2006), the virus under inspection (or parts of its genome engineered into a cultivatable reference strain) is exposed to different concentrations of a single drug, and its replication is measured. The result is a curve (Figure 2.7) that represents the decreasing activity of the virus with increasing drug concentrations. The most common summary measure for this curve is the 50% inhibitory concentration — that is, the drug concentration that halves viral activity. The fold-change in the 50% inhibitory concentration between a patient’s virus and a sensitive reference isolate, sometimes called the resistance factor,

$$\text{RF} = \frac{\text{IC}_{50}(\text{clinical virus})}{\text{IC}_{50}(\text{reference virus})}$$

is then used as a scalar representation of resistance to a specific drug. Commercial phenotyping is available, for example from the companies Virco<sup>6</sup> and Monogram Biosciences<sup>7</sup>.

## 2.6 Challenges and opportunities for modeling

One aim of this chapter was to review the necessary background on HIV. Another aim was to raise awareness for the opportunities and challenges to modeling in the field

<sup>6</sup>[www.vircolab.com](http://www.vircolab.com)

<sup>7</sup>[www.monogrambio.com](http://www.monogrambio.com)

of anti-HIV therapy. In particular, rational (or model-based) therapy would greatly benefit from advances in the following areas:

- Models linking variations in the viral genotype, differences in phenotypic (*in vitro*) properties, and individual response to therapy *in vivo*.
- Models tailored to the characteristics of specific novel drug classes.
- Models for viral evolution in the presence and absence of drug, and their exploitation in the clinical management of HIV infection.

All of these challenges will be addressed in this thesis. The research includes development of novel methodology, application to clinical data, comprehensive evaluation, and implementation in software. Throughout the thesis, particular attention is devoted to three modeling strategies:

- *Expanding* models to accommodate additional input data, such as host markers or clinical parameters.
- *Refining* models by incorporating domain-specific statistical, evolutionary, or structural aspects.
- *Interpreting* and *visualizing* models and data to discover novel knowledge, avoid “black-box” solutions, and encourage both clinical application and experimental validation.

The implementation of these strategies in specific problems depends heavily on the precise goals, the amount of data and the prior knowledge available. Often, several approaches may be feasible and priorities have to be set. These issues shall be discussed within the general context of modeling in the next chapter.



---

# Modeling with Data: The Views of Aristotle and Plato, and Later Developments<sup>1</sup>

Science *is* modeling. Scientific models are formal representations of natural phenomena that must be able to make quantitative predictions about unobserved events.<sup>2</sup> The work presented in this thesis spans three “cultures” of modeling: machine learning (e.g. Section 5.3), classical statistics (e.g. Section 4.2), and domain-specific explanatory modeling (e.g. Sections 4.3 or 6.4). The main purpose of this chapter is to provide a unifying perspective by showing how these cultures are concerned with the same problems, differing mainly in their adoption of more empirical or more explanatory perspectives.

## 3.1 Generality, predictive power, explanatory power

Three important, and widely discussed, characteristics of models are:

- *Generality*: the breadth of natural phenomena for which a model claims to apply.
- *Predictive power*: the degree to which a model agrees with observations.
- *Explanatory power*: the degree to which the architecture of the model reflects the actual mechanisms of nature.

Generality is a very intuitive concept. A more general model is usually regarded as superior from a scientific point of view. However, if the predictive power of a

---

<sup>1</sup>The title of this chapter is chosen in humble reference to (Lehmann, 1990).

<sup>2</sup>Many researchers and philosophers of science also require scientific models to be *falsifiable*. This means that the unobserved events for which predictions can be made need to be observable *in principle* (Popper, 1959).

less general model is sufficient for some phenomena, it may be preferred to a more general model in these cases (e.g. classical vs. quantum mechanics). Note that given two models, it need not be the case that one of them is more general than the other. Predictive power is also intuitive. Obviously, models with higher predictive power are preferred in general. However, often there is a trade-off between predictive power and the computational demands of a model. Thus, for many purposes, predictive power is deliberately sacrificed in favor of computational feasibility.<sup>3</sup> Heisenberg's uncertainty principle suggests that there are upper bounds to predictive power at least in some modeling domains.

Explanatory power, while superficially intuitive, is a more complex and controversial concept, as it involves assumptions on the relationship between models and reality.<sup>4</sup> Moreover, it is controversial whether explanatory power is desirable or even achievable (reviewed in Blackburn, 2005). Two extreme viewpoints are taken.

## 3.2 Scientific realism vs. instrumentalism

*Scientific realism* declares the goal of science “to find the real law of Nature” (Vapnik, 2005). For realism, scientific models are about “something *out there*, ‘external’ and (largely) independent of us”, about “an objective, external world” (Fine, 1986), and the elements of a model are equated with elements of the “real” world. Consequently, realism judges models by both their predictive *and* explanatory power. Scientific realism is often traced back to Aristotle (e.g. Duhem, 1969), who distinguished between “understanding the *fact* and understanding the reason *why*” (quoted from Freudenthal, 2003).

In contrast, *instrumentalism* (also called scientific anti-realism), considers models as “nothing but computation rules” (Popper, 1968), “a heuristic device, a calculating instrument for predictions alone” (Rosenberg, 2000, p.103), thus “denying that scientific theories seek to describe the underlying realities that systematize and explain observational generalizations” (Rosenberg, 2000, p.103). In language more familiar to our community, Nature is assumed to form “the outputs  $y$  from the inputs  $x$  by means of a black box” (Breiman, 2001b), and models will not aim to be literal descriptions of the inner working of this black box. Consequently, instrumentalism judges models by their predictive power, and possibly by their generality, but does not attempt to define criteria for explanatory power. Instrumentalism is often traced back to Plato (e.g. Duhem, 1969), who believed that “genuine scientific knowledge of our material world is impossible” (Good, 1999), and who “started the tradition of ‘saving the appearances’

---

<sup>3</sup>For example, Baker and Sali (2001) have compared different scenarios in which *de novo* protein structure prediction, threading, or comparative modeling might be most appropriate.

<sup>4</sup>In the philosophy of science, such assumptions are termed epistemological, or meta-theoretical, (cf. e.g. Rosenberg, 2000).

in astronomy” (Niiniluoto, 2007).

Ever since the times of Plato and Aristotle, discordant interpretations of models from realistic vs. instrumentalistic attitudes have been a source of controversy and heated debate. For example, the famous controversy between Galileo Galilei and the Catholic Church in 1633 – considered by many as the starting point of the “scientific revolution” (Kuhn, 1957) – was about whether the Copernican (heliocentric) model of planetary motion should be interpreted realistically (Galileo) or instrumentalistically (Church). While scientific realism appears to reflect the motivation of scientific endeavor most naturally, there are substantial concerns about the feasibility of such interpretations, which have led to claims that “the realist programme has degenerated by now to the point where it is quite beyond salvage” (Fine, 1986). Most notably, quantum mechanics, “the most precisely tested and most successful theory in the history of science” (Kleppner and Jackiw, 2000), is interpreted from a decidedly instrumentalistic point of view by most researchers, for example by denying the wave function a “real” existence.<sup>5</sup> A particularly striking example – in which instrumentalistic interpretation is inevitable – is given by attempts to provide models for effective procedures, or “algorithms”. Dozens of completely different models for effective procedures have been shown to be mathematically equivalent (e.g. Turing machines, lambda calculus, or  $\mu$ -recursive functions (Taylor, 1998)). To date, no “recipe” for performing a specific task has been described which could not be instantiated in any of these models. Thus, while all models are equivalent, and possibly “perfect” in the sense that they give a complete picture of effective computation, at most one of them could be “true” from a realistic point of view.

### 3.3 Explanatory vs. empirical models

As discussed above, from a realistic perspective, models are not only judged according to generality and predictive power, but also by their explanatory power. Models that merely aim to “save the appearances”, i.e. provide predictive power without an attempt to suggest a physical explanation of the modelled phenomena, are often termed *empirical* models. In contrast, models suggesting to mimic the underlying physical principles of a system are termed mechanistic, or *explanatory* models. Healy (1978) polemizes this dichotomy by distinguishing between “technological” vs. “scientific” models, stating that “in contrast to the scientist, the technologist is not concerned with truth at all.” (Healy, 1978).

Empirical models are widely and successfully used throughout science and technology. Their use can be advocated using pragmatic or instrumentalistic arguments. The pragmatic attitude contends that empirical models are “better than nothing” in

---

<sup>5</sup>This is part of the “Copenhagen” or “standard” interpretation of quantum mechanics first advocated by Bohr and Heisenberg (reviewed in Ismael, 2004)

situations when little is known about the system to be modelled, where there are numerical identifiability or parameterization problems, or where a more general or mechanistic model might be computationally intractable. Often, empirical components are incorporated within a generally mechanistic model, which is often referred to as a semi-mechanistic model, for example in quantum chemistry or in pharmacokinetic/dynamic modeling. From an instrumentalistic attitude explanation is not necessary (or even possible), leaving mainly generality and predictive power as criteria for judging models. Possibly the oldest justification for using empirical models is due to the stoic philosopher Geminus (c.10 BC – c.60 AD), who stated that “it is no part of the business of the astronomer to know what is by nature suited to a position of rest, and what sort of bodies are apt to move, but he introduces hypotheses under which some bodies remain fixed, while others move, and then considers to which hypotheses the phenomena actually observed in the heavens will correspond” (quoted from [O’Connor and Robertson, 2003](#))).

In Section 3.5.2, we shall discuss, using the example of planetary motion, that there is in fact not a strict dichotomy between empirical and explanatory models. Rather, all explanatory models rely on some empirical components or assumptions. Thus, instead of a dichotomy, it may be more appropriate to think of a continuum from more empirical to more explanatory models. For example, Newton, whose model of planetary motion explained the elliptical motion which was empirically assumed in Kepler’s model (cf. Section 3.5.2), was very much aware of the unexplained assumptions in his own model (most notably the nature of gravity), as apparent from his famous phrase: “I have not yet been able to discover the cause of these properties of gravity from phenomena and I feign no hypotheses” ([Cohen and Whitman, 1999](#), p.943).

### 3.4 Statistical vs. “machine learning” models

In this section, we shall briefly review a modeling discussion in the statistics community, which at its core is also about more empirical vs. more explanatory modeling. Central articles in this debate include ([Lehmann, 1990](#); [Chatfield, 1995](#); [Breiman, 2001b](#)), although the roots of the discussion go back to [Neyman \(1939\)](#), who distinguished between “explanatory” vs. “interpolatory” models. The classical approach to developing more empirical models has been statistics (in the frequentist, parametric paradigm): First, a stochastic data model for a phenomenon to be modelled is established. Then, an attempt is made to estimate the “true” values of the parameters from data about the phenomenon. The parameterized model can then be used for prediction or – from the perspective of scientific realism – in attempts to learn and understand more about the underlying process.

In this classical approach, it is implicitly assumed that enough domain knowledge exists that allows for pre-specifying the model accurately. This assumption has

been criticized widely, for example “as being at worst an arbitrary imposition of the modeler’s assumptions on the data and at best an inflexible approach to modeling data” (Jordan, 1994). The relatively late increase in interest on model specification could be due to the fact that it was considered from the beginning as “entirely a matter for the practical statistician” (Fisher, 1922). In other words, “the theoretician is happy to accept that his abstract probability triple  $(\Omega, A, P)$  was found under a gooseberry bush, while the applied statistician’s model ‘just grew’ [ . . . ]” (A. P. David, quoted in Lehmann, 1990).

While statistics considered itself the “home” of empirical modeling, a “second culture” (Breiman, 2001b) grew rapidly from outside statistics, defining its activity as “machine learning” or “data mining”. Machine learning might be defined as empirical modeling without imposing constraints on the stochastic nature of the data generation process. Rather, the data generation process is treated as unknown, as a “black box” (Breiman, 2001b). This paradigm is taken to the extreme in situations “where the analyst looks at a new set of data with virtually no preconceived ideas at all” (Chatfield, 1995). It is assumed that “the correct model is truly unknown” and that the models considered “need not be of the same form and none of them need to be correct” (Hosking et al., 1997). This opposition to classical parametric statistics is also reflected in a new terminology, in that models are “learned” instead of “parameterized”. In the mid- to late 1990s, the statistics community was surprised by the rapidly increasing popularity of this other culture (Chatfield, 1995; Hosking et al., 1997; Friedman, 1998) and it was realized: “We are no longer the only game in town. Until recently if one were interested in data analysis, Statistics was one of the very few even remotely appropriate fields in which to work. This is no longer the case. There are now many other exciting data oriented sciences that are competing with us for customers, students, jobs, and our own statisticians.” (Friedman, 1998).

Arguably, the roots of both cultures trace back to 1823, when Carl Friedrich Gauss (1777–1855) described his method of least squares. We shall use this foundational contribution as an example for coarsely outlining the different flavour of statistics and machine learning. Suppose we are given data in the form of a vector of samples  $(x_1, y_1), \dots, (x_n, y_n)$ , with  $x_i \in \mathbb{R}^{p-1}$  and  $y_i \in \mathbb{R}$ . Suppose also that our model for these data is  $f(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$ , i.e. a simple linear relationship between inputs and outputs. For notational convenience let  $y = (y_1, \dots, y_n)^T$ , and denote by  $X$  the  $n \times p$  matrix

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}.$$

A typical machine learning argumentation would be not to assume anything additionally about the data generation process or about the role and distribution of noise. Rather, “learning” would consist in finding a model that is optimal according to some loss function (e.g. [Schölkopf and Smola, 2002](#), p.61), most commonly the squared error loss. In this formulation the desired set of parameters  $\hat{\beta}$  would be specified as

$$\hat{\beta} = \arg \min_{\beta} (y - X\beta)^T (y - X\beta).$$

This formulation, which is closest to Gauss, is purely geometrical, and so far contains no additional assumptions about the data generating process and the noise.

In contrast, classical statistical modeling requires an explicit formulation of the data generating process, including the distribution of noise. This is accomplished by “constructing a hypothetical infinite population, of which the actual data are regarded as constituting a sample. The law of distribution of this hypothetical population is specified by [...] parameters.” ([Fisher, 1922](#)). The simplest approach is to assume that samples are drawn independently from each other, and that they are equipped with noise that is normally distributed. More formally, the data model would be

$$y = X\beta + \epsilon,$$

with

$$\epsilon \propto \mathcal{N}(0, \sigma^2 I).$$

The search criterion is then also defined statistically, for example in terms of the likelihood function  $L_x(\beta) = P_{\beta}(x)$ . Then the desired parameters are those that maximize the likelihood function:

$$\hat{\beta} := \arg \max_{\beta} L_x(\beta).$$

It is easy to see that the parameter vector that maximizes the likelihood is also the one that minimizes the sum of squares. Thus, in this example, parameter estimation and “learning” will lead to the same fitted model.

Examples of approaches originating from the “second culture”, where the necessity for modeling the data generation process is abandoned, include nearest neighbor methods (reviewed in [Hastie et al., 2001](#), pp. 411–436), decision trees (reviewed in [Hastie et al., 2001](#), pp.266–279), rule set induction (reviewed in [Mitchell, 1997](#), pp. 274–306), neural networks (reviewed in [Bishop, 2006](#), pp. 225–284), support vector machines (reviewed in [Schölkopf and Smola, 2002](#)), or ensemble methods such as bagging or boosting (reviewed in [Hastie et al., 2001](#), pp. 299–344), or random forests ([Breiman, 2001a](#)). It should be noted that ensemble methods and other extreme forms of generic empirical modeling pose serious challenges to the interpretability of the learned model: “Doctors can interpret logistic regression. There is no way they can interpret a black box containing fifty trees hooked together.” ([Breiman, 2001b](#)).

Even though some of these models can be interpreted from the point of view of statistical modeling (e.g. (Jordan, 1994, decision trees), (Pontil et al., 2000, SVMs), (Friedman et al., 2000, Boosting)), in most cases, the reformulations are less elegant and less general than the original descriptions of the methods. While some methods, such as SVMs or neural networks have in common with statistical models that the method is based on an objective function which is optimized during learning, for other methods (e.g. trees, random forests, boosting) there is not even an objective function in the original formulation, and “the algorithm is the true starting point” (Faraway, 2006).

Often, explicitly modeling the data generation and noise processes will be more explanatory – if successful – than the use of very generic machine learning methods. Thus, developing a statistical model may be appropriate when more domain knowledge is available. In contrast, machine learning (or data mining, or nonparametric regression) is especially suited for situations in which very little known about the phenomenon to be modelled, and for complicated domains with many variables of uncertain importance. Consequently, the approaches developed in these fields rely on weaker assumptions as those in classical parametric statistics. Still, there is no learning method without “inductive bias” (e.g. Mitchell, 1997, pp. 39–45). For example, in decision tree learning, it is assumed that the predictor space can be adequately partitioned into axis-parallel regions. The famous “No Free Lunch” theorems (reviewed in Wolpert, 2001) imply (among many other things) that no learning method can outperform any other method *in general*.

As in the discussion of empirical vs. mechanistic models in the previous section, we suggest to consider “statistical” vs. “machine learning” modeling not as a dichotomy, but rather as a continuum in the range of more empirical models, extending from empirical models with some explanatory components to highly generic empirical models. This “continuum” view is supported by several observations. Firstly, there are many connections between partially explanatory statistical modeling (which would generally still be regarded as part of empirical methodology) and modeling approaches that reflect genuinely mechanistic or explanatory thinking. For example, the “population-based” approach to pharmacokinetics is often based on a combination of mechanistic, compartmental pharmacokinetic or pharmacodynamic modeling using differential equations with an estimation procedure based on nonlinear mixed-effects modeling (Pinheiro and Bates, 2000). Support for the “continuum” perspective is also provided by the increasing “cross-talk” between the disciplines. For example, more recent textbooks in applied statistics devote entire chapters to concepts and methods originating in machine learning (e.g. Wasserman, 2004; Faraway, 2006). Using labels such as “greater statistics” (Chambers, 1993); “wide view of statistics” (Wild, 1994), or “learning from data” (Hastie et al., 2001), statistical methodologists have been advertising the expansion of statistical modeling to encompass the whole modeling process, including model specification and selection. In turn, statistical methods are becoming more and more popular in machine learning, and much work is devoted to develop-

ing statistical foundations for the field (e.g. [Vapnik, 1998](#)). The defining difference between this paradigm of “statistical learning”, as compared to traditional parametric statistics is that the underlying distribution class is considered unknown. Thus, theorems should usually hold for *any* distribution on  $\mathcal{X} \times \mathcal{Y}$ . Famous examples include Vapnik-Chervonenkis type inequalities or other concentration of measure inequalities ([Schölkopf and Smola, 2002](#), chapter 12). The confluence of ideas from machine learning and classical statistics is also apparent in fields such as *semi-parametric* regression (e.g. [Ruppert et al., 2005](#)), which explicitly seeks to develop models combining parametric with nonparametric components.

Ultimately, “any method of analysis should be [...] judged on whether it successfully predicts or explains something. Statistical models may achieve this, but algorithmically based methods are also competitive” ([Faraway, 2006](#)). Depending on the problem at hand and the amount of domain knowledge available, in empirical modeling, “the best solution could be an algorithmic model, or maybe a data model, or maybe a combination.” ([Breiman, 2001b](#))

## 3.5 Examples

In this section, we try to substantiate the abstract discussion of the previous sections with some illustrative examples.

### 3.5.1 Empirical models and the “Rashomon” effect

We begin with an example given by [Breiman \(2001b\)](#). Consider the following three alternative linear models:

$$\begin{aligned} y &= 2.1 + 3.8x_3 - 0.6x_8 + 83.2x_{12} - 2.1x_{17} + 3.2x_{27} \\ y &= -8.9 + 4.6x_5 + 0.01x_6 + 12.0x_{15} + 17.5x_{21} + 0.2x_{22} \\ y &= -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 + 3.4x_{11} + 7.2x_{28} \end{aligned}$$

These are three five-variable linear regressions selected from 30 variables that are all within 1.0% test set error of each other. The example exhibits what Breiman calls the “Rashomon effect”: “that there is often a multitude of different descriptions [equations  $f(x)$ ] in a class of functions giving about the same minimum error rate” ([Breiman, 2001b](#)). From an instrumentalist perspective, this is not a source of concern, as long as the predictive power of these models is sufficient (evaluated on independent test data).

However, from the point of view of scientific realism, where we are not only concerned about predictive, but also about explanatory power, having to choose among these models is a disturbing task. Obviously, none of the three models describes “the real laws of Nature” (Section 3.2) — rather, they are purely empirical (Section 3.3),

i.e. chosen to fit the data (or, an independent validation set) as well as possible, without regard to their relation to true underlying processes.

### 3.5.2 Modeling planetary motion

In this section we shall review *the* classical example of scientific modeling: planetary motion. The example will underline the statement from Section 3.3 that empirical vs. explanatory modeling should be regarded not as a dichotomy, but rather as a continuum of possibilities.

For an observer on Earth, the trajectories of the planets appear very complicated, and it was considered a major challenge since Ancient Greece to accurately predict (and possibly, but not necessarily, explain) these trajectories. One particularly bewildering phenomenon was that of *retrograde motion* – a change in direction of a planet over a time frame of several weeks when observed from the Earth (Figure 3.1). We now know that the complexity of the observed trajectories is due to the revolution of the planets around the Sun, combined with the rotation of the Earth around itself (even complicated by a tilted rotation axis, i.e. one which is not orthogonal to the plane of revolution).

A model developed by Ptolemy (c.90–c.168 AD) became the basis of all further work for nearly 1,500 years. In this model, the Earth is located at the center. The Sun and the other planets revolve around this center in circular orbits. However, the movement of a planet does not take place on the cycle itself, but rather around an *epicycle*, another cycle of smaller radius which is moving itself along the main cycle (Figure 3.1(a)). The Ptolemaic model was famously challenged in 1543 by Nicolaus Copernicus (1473–1543). In his treatise “On the Revolution of the Celestial Orbs”, Copernicus proposed a model of circular planetary motion with the Sun, rather than the Earth, at the center<sup>6</sup> (Figure 3.1(b)), sharing however with the Ptolemaic model the use of epicycles.

The Copernican model, despite its huge explanatory implications, did not fit the data better than the Ptolemaic model. In this bewildering situation of many conceptually and computationally different models, Tycho Brahe (1546–1601), aged only 17, formulated a life-long research agenda in his diary: “I’ve studied all available charts of the planets and stars and none of them match the others. There are just as many measurements and methods as there are astronomers and all of them disagree. What’s needed is a long term project with the aim of mapping the heavens conducted from a single location over a period of several years.”<sup>7</sup> This agenda led to the first systematic large-scale data collection effort in the history of science: Between the years 1582 and

<sup>6</sup>While heliocentric models had been proposed already in ancient India and Greece, they gained little popularity at that time.

<sup>7</sup>Quoted from Wikipedia, [http://en.wikipedia.org/wiki/Tycho\\_Brahe](http://en.wikipedia.org/wiki/Tycho_Brahe), version: May 13, 2007.

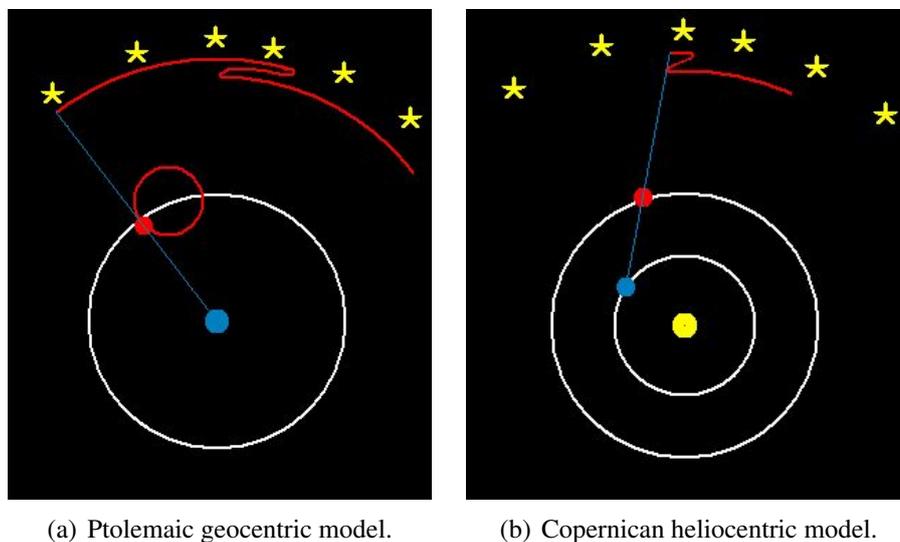


Figure 3.1: Heliocentric vs. geocentric models of planetary motion (from <http://faculty.fullerton.edu/cmconnell/Planets.html>). The phenomenon of apparent retrograde motion can be seen in the planet trajectory projected onto the celestial sphere (red lines).

1600 Brahe collected, among other data, a series of 923 measurements of the declination of the planet Mars. With Tycho’s death, Johannes Kepler (1571–1630), originally an assistant of Tycho, began to analyze the data on his own. After over six years of work, thousands of pages of calculations, and many different attempts and dead ends (Koestler, 1989), he came to the conclusion that all previously suggested models were insufficient in explaining Brahe’s data up to the limits of measurement error. Moreover, he proposed a new model in which Mars (as well as the other planets) moves in elliptical orbits with the Sun at a focus.

From a modern perspective it should be noted that the proposed heliocentric models (with or without epicycles) can all be described by means of an observationally equivalent geocentric model with epicycles (Hanson, 1960; Gearhart, 1985; Fitzpatrick, 2006). In fact, “there is no [...] curve used in any branch of astrophysics or observational astronomy today which could not be smoothly plotted as the resultant motion of a point turning within a constellation of epicycles, finite in number [...]” (Hanson, 1960). Thus, from an instrumentalist perspective, the use of a geocentric reference would be simply regarded as a change in the coordinate system for practical purposes, without any physical implications. In terms of predictive power, there is no advantage at all to the heliocentric models of Copernicus or Kepler, as compared to geocentric models in the spirit of Ptolemy. Rather, the difference between heliocentric and geocentric modeling lies in their explanatory power: in the geocentric model, retrograde

motion is predicted accurately by assuming that planets move on rather complicated epicyclical trajectories. In contrast, in the heliocentric model of Kepler, planetary motion proceeds along much simpler (elliptical) trajectories, and retrograde motion is *explained* as only perceived and apparent for an observer on Earth, rather than “real”. In conclusion, geocentric and heliocentric models stand side by side as computational tools, but heliocentrism offers a much simpler explanation of retrograde motion. As mentioned in Section 3.2, the famous controversy between Galilei and the Catholic Church in 1633 was precisely about whether heliocentrism should be interpreted from the perspective of instrumentalism or from that of scientific realism.

While Kepler’s model explained retrograde motion as a consequence of observing an elliptical motion from a moving planet, it did not provide an explanation for the cause of elliptical motion: the model class of ellipses was assumed on a purely empirical basis. Isaac Newton (1643–1727), in the third volume of his *Principia*, described a model of planetary motion which was equivalent in predictive power to Kepler’s model, but which far superseded it in terms of generality and explanatory power. In Newton’s model, Kepler’s empirically derived model of elliptic orbits appears as a special case from a much more comprehensive model for predicting the motion of all bodies based on the concept of forces acting between the bodies according to three axioms, with the forces themselves originating from another axiom, the law of universal gravitation.

Newton’s work provided a mechanistic explanation for Kepler’s model with observational equivalence. However, like its predecessors, this model is not devoid of empirical elements and unexplained concepts. Most importantly, there is no explanation of how gravitational forces should be mediated, and how this should happen instantaneously (particularly, at faster than light speed). Newton himself was aware of this explanatory inadequacy, but tried to put it into perspective with an instrumentalistic argumentation: “It is enough that gravity does really exist and acts according to the laws I have explained, and that it abundantly serves to account for all the motions of celestial bodies.”<sup>8</sup> In the 20th century, Albert Einstein resolved this action-at-a-distance problem in his general theory of relativity by introducing a curved spacetime. However, Einstein’s model does in turn not explain the mechanisms by which spacetime becomes curved by mass and energy.

Thus, the history of models for planetary motion appears as a history of providing explanations for previously empirical assumptions: Kepler’s model explained the retrograde motion unexplained in Ptolemy’s model, Newton’s model explained the elliptical trajectories unexplained in Kepler’s model, and Einstein’s model explained the apparent action-at-a-distance of gravity unexplained in Newton’s model. We see how explanatory improvements in a sequence of models were achieved by subsequently introducing more general, or “fundamental” concepts. However, these more general

---

<sup>8</sup>Quoted from Wikipedia’s article on “Newton’s law of universal gravitation” (version: April 22, 2007), which in turn quotes from (Westfall, 1978).

concepts are then left as empirical assumptions without further explanation.

### 3.5.3 Modeling sunflower growth

This is one of two examples in which we consider how more explanatory models can lead to improvements in predictive power, as compared to purely empirical models — if the structure of the phenomenon assumed in the explanatory model matches reality sufficiently well.

Reed and Holland (1919) have collected weekly measurements of the height of 58 sunflowers grown under uniform conditions.<sup>9</sup> Growth of plants or animals provides a very simple situation in which *a priori* background knowledge is available from observations of many different species: growth begins at a slow rate, becomes increasingly fast over time, but finally slows down until it eventually comes to a halt.<sup>10</sup> Moreover, the height of a plant or animal can be assumed to never decrease. A classical model for such a pattern is the logistic model,

$$f(t) = \frac{\phi_1}{1 + e^{-\frac{t-\phi_2}{\phi_3}}}.$$

Figure 3.2 shows the mean heights of the 58 sunflowers at weekly intervals between 7 and 84 days. Only the heights at 28, 35, 42, 49, and 56 days (shown in red) are used for model fitting. Predictions are then made for the entire time span. Models used are empirical approaches based on linear regression (dashed), polynomial (order-four) regression (dotted), support vector regression with radial basis function kernel and standard hyperparameters (dash-dotted), and a less empirical model based on the assumption of logistic growth (solid). It is obvious that all three empirical models lead to unrealistic predictions: linear regression predicts indefinite growth, fourth-order polynomial regression predicts shrinkage after a period of growth, and support vector regression predicts a shrinkage-growth-shrinkage pattern. Only a less empirical model, incorporating the background knowledge that (a) no shrinkage will occur, and (b) growth will ultimately slow down and come to a halt, leads to qualitatively reasonable predictions.

### 3.5.4 Modeling in population genetics

The *sunflower* example introduced above showed how a slightly less empirical model (incorporating some background or *a priori* knowledge on plant growth) could greatly improve over very empirical, off-the-shelf methods. We now turn to an even more

<sup>9</sup>This publication was identified by searching on the Web with terms such as “logistic”, “bilogistic”, and “growth data”.

<sup>10</sup>Of course there are indefinitely many other growth patterns and corresponding models, which are outside the scope of this thesis, cf. (Case, 2000).

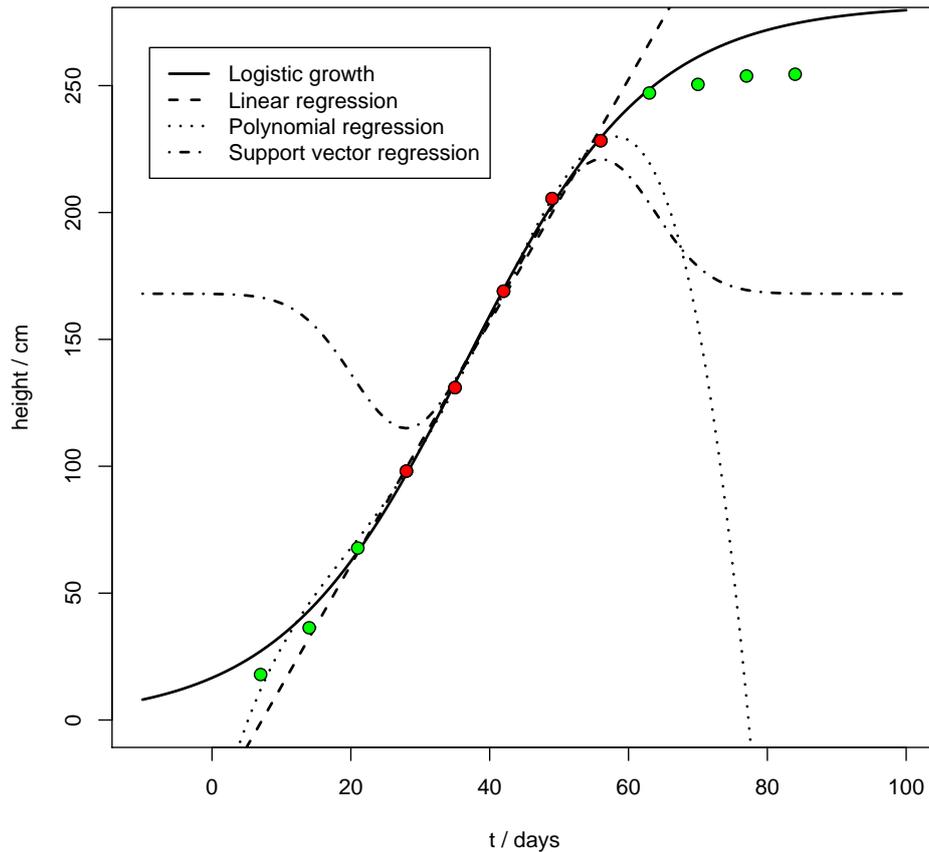


Figure 3.2: Sunflower growth measurements (points) reported in (Reed and Holland, 1919). The horizontal axis represents the time in days after onset of the study, the vertical axis represents the average height of the considered sunflowers in centimeters. The lines represent three different models fitted to the red points: a logistic growth model fitted by nonlinear least-squares minimization (solid line), a fourth-order polynomial regression model (dashed line), and a support vector regression model (dotted line). The green points are used as a test set to estimate the generalization ability of the methods.

extreme example, in which we show the differences in predictive performance (and also interpretability) that can be gained when the nature of the data-generating process is modelled correctly. The system is a simplified form of that studied in Section 4.1. Here, we consider an infinite population whose individuals consist of two biallelic (0 or 1) genetic loci. Thus, the state of the population can be described by the fraction of genotypes:  $p_{00}$ ,  $p_{01}$ ,  $p_{10}$ , and  $p_{11}$ , per definition summing up to one. Population dynamics is introduced via operators for mutation and selection. Mutation is assumed to take place independently at each locus with a symmetric, constant mutation rate of  $\mu$ , leading to the mutation matrix shown in Section 4.1.3. Here, we assume a fixed “weak” selection, with the following fitness values  $w_g$  assigned to genotypes  $g$ :  $w_{00} = 1$ ,  $w_{01} = 1.01$ ,  $w_{10} = 1$ ,  $w_{11} = 0.99$ . Thus, the system obeys the following dynamics (explained in detail in Section 4.1.3):

$$p'(t) = \frac{\langle w, Mp(t) \rangle}{\| \langle w, Mp(t) \rangle \|} - p(t)$$

We simulate 1,000 trajectories with random initial genotype frequencies (uniformly drawn from the population simplex, cf. Section 4.1.3), and random mutation rate (uniformly distributed in  $[0, 1]$ ), but always the same fixed fitness values. These trajectories are sampled at times  $0, 1, \dots, 9$ . Our goal is to learn how to estimate the mutation rate from observed trajectories.

Gaussian noise with mean zero and standard deviation  $\sigma$  is added to the trajectories. After adding noise, genotype frequencies at each time point are renormalized to one. The 1,000 simulations are replicated with the standard deviations

$$\sigma \in \{0.0, 0.001, 0.0025, 0.005, 0.0075, 0.01, 0.05, 0.1\}.$$

An example trajectory with  $\sigma = 0.01$  is shown in Figure 3.3.

We first adopt a completely agnostic approach to predicting mutation rates from population trajectories. We use support vector regression with radial basis function kernel (insensitive loss is set to  $\epsilon = 0.01$ , otherwise standard parameters are used) to derive models from trajectories with known mutation rates. Ten-fold cross-validation is used on the 1,000 trajectories with corresponding mutation rates to obtain unbiased estimates of predictive performance.

As an alternative, we assume that the form of the model is known, except for mutation rates. Using this model, learning from observations is not necessary. Rather, the mutation rate is predicted by minimizing the sum of squared differences between the noisy observations and simulations of the system with various mutation rates.

The predictive performance of these two different models is summarized in Table 3.1. As expected, the reliability decreases monotonically with increasing noise. In general, the mechanistic model far outperforms the empirical model. However, the performance decrease between noise rates  $\sigma = 0.01$  and  $\sigma = 0.1$  is much sharper for

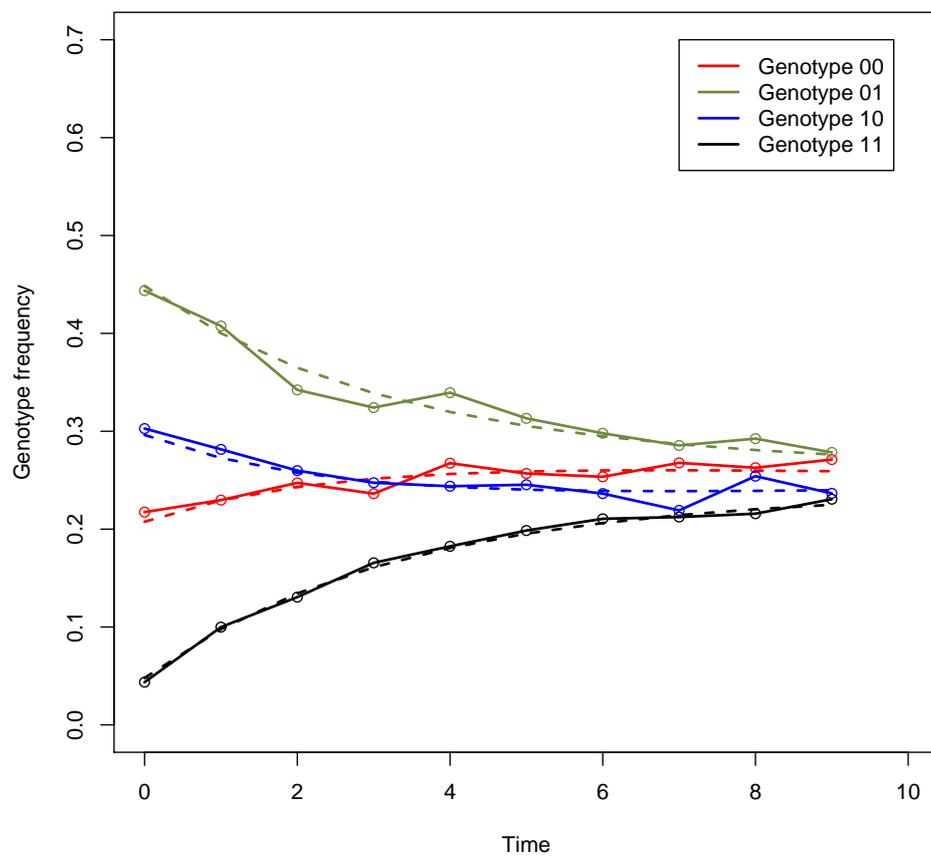


Figure 3.3: Sample trajectory of the two-locus system of population dynamics. The dotted lines represent the numerical solution to the system, while the solid line connects the samples taken at time points 0 to 9 with Gaussian noise of mean zero and standard deviation 0.01 added.

the mechanistic model. At a noise of 0.1 the empirical model even slightly outperforms the mechanistic model, although both predictions are highly unreliable in that case, since noise has become the dominating influence in the observations.

Noise	MAE		$r^2$	
	SVM	ODE	SVM	ODE
0.0	0.11	0.00	0.71	1.00
0.001	0.11	0.00	0.71	1.00
0.0025	0.11	0.01	0.69	0.99
0.005	0.12	0.02	0.68	0.98
0.0075	0.12	0.03	0.67	0.96
0.01	0.13	0.04	0.63	0.93
0.05	0.19	0.17	0.34	0.40
0.1	0.22	0.25	0.15	0.13

Table 3.1: Predictive performance of support vector regression (SVM) and non-linear least squares fitting to the mechanistic model (ODE). The comparison shows mean absolute error (MAE) and squared correlation coefficient ( $r^2$ ) between predicted and true mutation rates for various magnitudes of noise in the observations.

### 3.5.5 Summary

The previous examples have underlined the following points:

- Explanatory modeling aims to be faithful to what we think are the true mechanisms of a natural process. Explanatory and empirical modeling are not to be seen as a dichotomy, because each model includes certain empirical components. From an antirealistic point of view, the explanatory power of a model is not a relevant characteristic.
- A model that is “wrong” from a realistic point of view may be observationally equivalent (i.e. behave identically in any respect) to a more “correct” model. This is the case for geocentric epicycle models in the tradition of Ptolemy which are observationally equivalent to the later heliocentric models.
- Empirical models based on few assumptions can have drastically higher predictive performance than explanatory models based on wrong assumptions. This indicates that statistical learning “can be applied to some problems where classical methods cannot be used” (Vapnik, 2005). For example, in an engineering application, Cejudo et al. (2002) concluded: “The neural network models are more

adequate to simulate wheel performance because, in this case, the hypotheses of the physical model are not in accordance with the reality.”

- If the assumptions of an explanatory model are appropriate, it can vastly outperform generic empirical models in terms of both predictive and explanatory power, as has been shown in the examples on sunflower growth and on population genetics.

## 3.6 Modeling strategies in this thesis

The level of explanatory power provided in this thesis is generally more limited than in biomedical fields with a more established tradition of explanatory modeling, for example pharmacokinetics and pharmacodynamics. In fact, many studies on mechanistic modeling in the context of HIV dynamics have been published, but they often rely on unrealistic assumptions regarding the genetic basis of drug resistance. Moreover, validations on real, independent data are scarce. In such a situation, the use of an inadequate mechanistic model would mean sacrificing predictive power compared to that achievable by modern statistical learning methods.

Figure 3.4 may serve as a metaphor for the current state of model-based anti-HIV therapy. It shows the original Mars data of Tycho Brahe<sup>11</sup> (black points) with a superimposed fit obtained using a modern empirical regression method (blue line). While the fit might be regarded as reasonable even in cross-validation (red points), it is obvious that the level of explanatory power is far below any of the models reviewed in Section 3.5.2, as there is not even a notion of moving bodies incorporated in the model. Still, in terms of predictive power, such a model would be far superior to a very inadequate mechanistic model based on, say, rectangular orbits.

Despite this primary focus on predictivity (rather than explanation) throughout this thesis, we describe several steps away from purely empirical modeling to the incorporation of domain-specific or semi-mechanistic elements. However, to ensure that these extensions are not seen as an exercise for its own sake, they are always benchmarked against purely empirical approaches. Whenever possible, the impact of variables (e.g. mutations or clinical parameters) affecting a modelled system is quantified and interpreted. Domain-specific background knowledge is incorporated in several places, for example in Section 4.3, in which a model of viral evolution is coupled with a method for predicting drug resistance via a custom Fisher kernel. Another example, reported in Section 6.4, is a model that explains changes in viral coreceptor usage phenotype in terms of changes in the side-chain orientation pattern of the viral envelope protein, thus taking into account the fact that phenotypic changes are ultimately mediated by physico-chemical alterations in specific proteins. It is not always straightforward to

<sup>11</sup>Retrieved from <http://www.pafko.com/tycho/observe.html>.

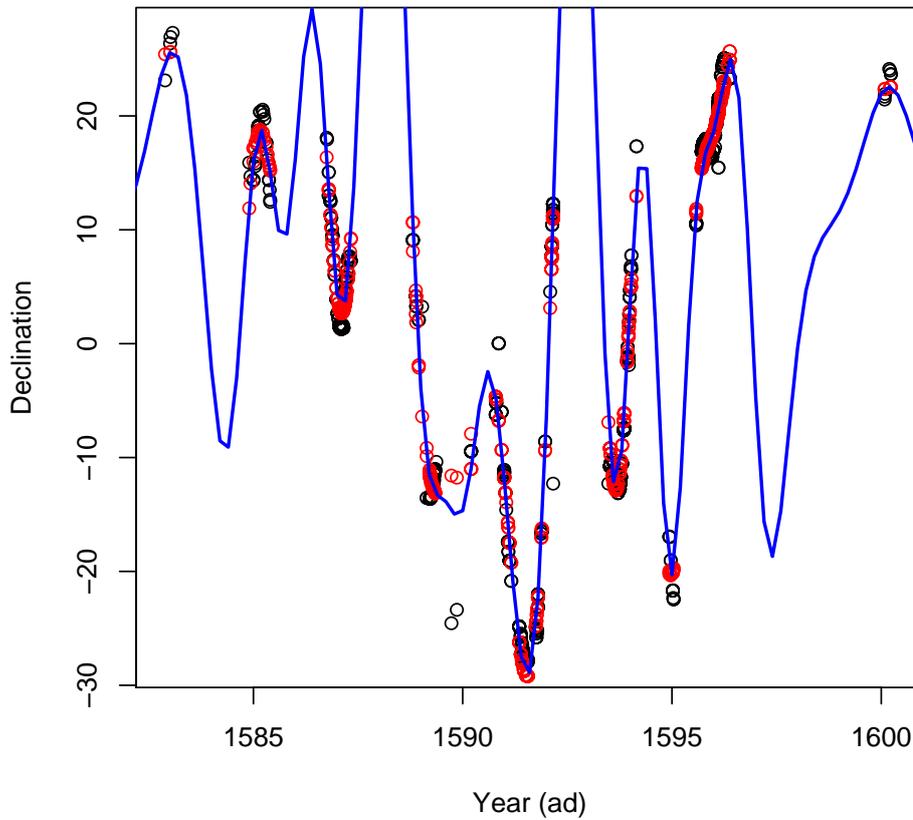


Figure 3.4: The data collected by Tycho Brahe, consisting of 923 observations recorded between the years 1582 and 1600. The observed values (shown as black points) show the declination of the planet Mars. Red points show predictions for these data using support vector regression from ten-fold cross-validation. The blue curve shows the fit of a support vector regression when using the entire data set.

decide on a particular modeling strategy *a priori*. Often, several alternatives are conceivable, and it is far from clear which one will be the most promising. For example, our first approach to the problem of modeling mutation loss during interruption of therapy was based on a linear mixed-effects model. However, the survival analysis approach presented in Section 4.2 proved to be superior both in terms of model fit and in terms of explanatory power.

In 1900, David Hilbert proclaimed the axiomatization of physics to be a major aim of 20th-century mathematics, a research programme which became known as number six in the famous list of Hilbert's problems (Schirmacher, 2003). His explicit goal was to win mathematicians "over to the inspirations that flow into mathematics from the side of physics" (quoted from Schirmacher, 2003). Without doubt, the subject of physics was chosen as an example because it was the most advanced among the sciences at that time, and because it had the most well-established tradition of quantitative modeling. However, Hilbert did not think of physics alone, rather he was deeply convinced that "any science is at any time not only ripe enough but necessarily requires axiomatization" (quoted from Schirmacher, 2003). One hundred years later, mathematics is at the verge of becoming "biology's next microscope (only better)" and in turn biology "mathematics' next physics (only better)" (Cohen, 2004). Is the time ripe for formulating a sixth Hilbert problem for biology and medicine? How would young paradigms such as data mining, machine learning, or hypothesis-free science fit into such an agenda? In the spirit of this chapter, we should praise them as opportunities for starting empirical modeling earlier than ever before, even in the complete absence of problem-specific knowledge. Ultimately, however, they are nothing but initial steps on the quest for a better understanding and explanation of Nature.



---

# Viral Evolution: Foundations and Clinical Implications

[Futuyma \(1998, p.4\)](#) defines biological evolution as “change in the properties of populations of organisms, or groups of such populations, over the course of generations”. He emphasizes that “the changes in populations that are considered evolutionary are those that are ‘heritable’ via the genetic material from one generation to the next”. Because many biological properties can be understood more easily in the light of evolution, evolutionary theory has been called “the one theory that transcends all of biology” ([Nowak, 2006](#)). It is important to understand that evolution happens on the level of populations, not on the level of an individual. In fact, we shall focus on the evolution of a population consisting of a single species; the study of interactions between groups of populations is part of the field of ecology ([Case, 2000](#)).

[Beerenwinkel \(2004, p.37–42 & p.55–69\)](#) provides a very good introduction to evolutionary theory. In addition, much of the relevant work is now conveniently available in standard text books. For these reasons, we shall only provide the minimal background necessary to understand the research presented in this chapter. This background will be introduced in the appropriate sections. Two general introductions to evolutionary theory, which are concise, highly readable, and complement each other well, are given by [Gillespie \(2004, focusing on classical population genetics\)](#) and [Nowak \(2006, focusing on evolutionary theory \*beyond\* population genetics\)](#). Specific monographs on HIV-1 modeling (mostly focusing on evolution and epidemiology) are also available ([Crandall, 1999](#); [Nowak and May, 2000](#); [Rodrigo and Learn, 2001](#); [Tan and Wu, 2005](#)). Methodologically, population genetic modeling is in no way different from any other kind of modeling in biology. General introductions to this broader field include ([Edelstein-Keshet, 2004](#); [Britton, 2003](#); [Allman and Rhodes, 2004](#); [Haefner, 2005](#); [Murray, 2002, 2004](#)).

This chapter is structured as follows. In Section 4.1, we study fitness landscapes and their impact on the evolutionary dynamics of a population. In the other two sec-

tions of this chapter, we ignore population dynamics, focusing exclusively on the dynamics of mutation fixation or loss within a population. Specifically, Section 4.2 is concerned with viral evolution in the absence of drug pressure. While treatment interruptions (TIs) are not a recommended treatment strategy against HIV, unplanned TIs are an inevitable part of antiretroviral therapy for many patients (e.g. due to non-adherence or suboptimal drug levels). To our knowledge, this is the first study on mutation dynamics based on methods from survival analysis. The final section (4.3) serves as a bridge from the present chapter to Chapter 5, since we show how to use evolutionary modeling for a clinical purpose, namely to improve the prediction of drug resistance from genotype. We build on ideas originating in differential geometric approaches to statistics and refined in the field of statistical learning to derive a so-called “Fisher kernel” for the mutation dynamics model of mutagenetic tree mixtures.

## 4.1 Shapes of fitness landscapes and population dynamics<sup>1</sup>

We study the relation between the shape of a fitness landscape and the fate of a population of multilocus genotypes, using the time to equilibrium as a proxy of evolutionary outcome. Shapes are shown to be related via symmetries determined by the genetic operators that govern the dynamics of the population. Using large-scale numerical simulations, we show that the shape of the fitness landscape is significantly associated with time to equilibrium. Recombination speeds up the onset of equilibrium within all shapes, although the magnitude of the effects appears to be shape-dependent. Classification of fitness landscapes according to their shape is shown to be complementary to a rank-based classification, suggesting potential synergies from a hybrid classification scheme.

### 4.1.1 Introduction

Fitness landscapes quantify the reproductive success of the individuals in a genetically heterogeneous population. They enter evolutionary models via a selection operator. Intuitively, the concept of fitness aims at providing a simple scalar-valued summary of the many different factors in the life cycle of organisms that affect their reproductive success, such as viability (probability to survive to reproductive age), mating success, age of reproduction, or fertility (number of produced offspring). Models that integrate selection with other genetic operators, such as mutation or recombination, usually lead to complex, nonlinear systems with analytically intractable dynamics. As a consequence, considerable attention has been devoted to the identification of features of

---

<sup>1</sup>The work reported in this section was performed in collaboration with Niko Beerenwinkel (Berkeley).

fitness landscapes that are associated with certain evolutionary “outcomes”.

One of the most widely studied features of fitness landscapes is *epistasis* (reviewed in Cordell, 2002), which quantifies fitness interactions between pairs of alleles at different loci. Epistasis has been linked to evolutionary properties of a system in various ways, most prominently in studies investigating the advantage of sex and recombination (e.g. Feldman et al., 1974, 1996; Grote et al., 1998; Weinreich, 2005). Epistatic effects in HIV-1 have also been subject to a large-scale study, with evidence for positive epistasis (Bonhoeffer et al., 2004).

Given its definition in terms of pairs of loci, classical epistasis may fail to discover or describe higher-order dependencies among loci in multi-allelic ( $n > 2$ ) scenarios. Addressing this shortcoming, Beerenwinkel et al. (2006) have suggested a novel classification system for fitness landscapes. In this system, landscapes are classified according to their geometric “shape”. In the case of two loci, the shapes of a fitness landscape correspond exactly to the different forms of classical epistatic interaction (positive, negative, no interaction). Based on this observation, the novel classification scheme has been hypothesized to play a role of similar importance in describing the evolution of multi-locus systems as classical epistasis has been found to play for biallelic systems. In this chapter, we provide empirical support for this hypothesis. For the sake of clarity and simplicity, we focus on a minimal extension of the classical case: a three-locus monoallelic system, whose dynamics is determined by infinite population size, continuous time, exclusively fitness-dependent selection, and locus-independent mutation and recombination rates. Moreover, we focus on only one proxy of the many possible measures of evolutionary outcome, namely the time until the system reaches equilibrium. We find a strong statistical association between the shape of the fitness landscape and the time to equilibrium.

### 4.1.2 The shape of a fitness landscape

In this section we briefly review the concept of the shape of a fitness landscape, as proposed in (Beerenwinkel et al., 2006).<sup>2</sup>

Genotype space and fitness landscape.

We consider the *genotype space*  $\mathcal{G} = \{0, 1\}^n$  (for most of the discussion we shall focus on the case  $n = 3$ ). A genotype  $g \in \mathcal{G}$  with  $g_i = 0$  is said to carry the *wild type* allele at locus  $i$ , whereas  $g_i = 1$  is referred to as the *mutant* allele. A fitness landscape on  $\mathcal{G}$  is a function  $w : \mathcal{G} \rightarrow \mathbb{R}_{\geq 0}$ . Each coordinate  $w_g$  of  $w$  denotes the logarithm of the reproductive fitness of genotype  $g$ . In order to classify landscapes according to their “shape” or “curvature”, one needs a continuous object, rather than a discrete assignment of fitness values to individual genotypes. The solution suggested

<sup>2</sup>We sacrifice some generality for ease of exposition.

by [Beerenwinkel et al. \(2006\)](#) is to extend  $w$  from individual genotypes to populations.

### Population simplex and genotope.

Here, we shall focus on populations of infinite size, such that a population can be considered a point in the *population simplex*  $\Delta_{\mathcal{G}} = \{p \in [0, 1]^{\mathcal{G}} : \sum_{g \in \mathcal{G}} p_g = 1\}$ . The cumulative allele frequencies in a population  $p$  are summarized as a point in the  $n$ -cube  $\Pi_{\mathcal{G}} = [0, 1]^n$ . More precisely, if  $v = (v_1, \dots, v_n) \in \Pi_{\mathcal{G}}$ , then  $v_i = \sum_{\{g \in \mathcal{G} : g_i = 1\}} p_g$ . Thus,  $v_i$  denotes the frequency of the mutant allele at locus  $i$  in the population  $p \in \Delta_{\mathcal{G}}$ . We refer to  $\Pi_{\mathcal{G}}$  as the *genotope*. Note that [Beerenwinkel et al. \(2006\)](#) consider a more general setting, in which the genotope need not be an  $n$ -cube. Rather, it can be any convex polyhedron, depending on the genotype space  $\mathcal{G}$ .

### Continuous fitness landscape.

The *continuous fitness landscape*  $\tilde{w}$  is defined as a scalar function on the genotope  $\Pi_{\mathcal{G}}$ , based on a fitness landscape  $w$ . The continuous fitness landscape maps allele frequencies to population fitness values. The fitness of a population is defined as the frequency-weighted average fitness of the individuals in the population:  $\langle p, w \rangle = \sum_{g \in \mathcal{G}} w_g p_g$ . A point  $v$  in the genotope (representing a specific allele frequency) corresponds to a fiber in the population simplex, consisting of all the populations with a given allele frequency. [Beerenwinkel et al. \(2006\)](#) define the value  $\tilde{w}(v)$  of the continuous fitness landscape at this allele frequency as  $\tilde{w}(v) := \max \langle p, w \rangle$ , where the maximum is taken over all populations  $p$  with allele frequency  $v$ . Obviously,  $\tilde{w}$  has the same values as  $w$  on the vertices of  $\Pi_{\mathcal{G}}$  (for  $w$ , these vertices are interpreted as single populations rather than allele frequencies).

### Shape of a fitness landscape.

The central insight underlying the novel classification of fitness landscapes laid out by [Beerenwinkel et al. \(2006\)](#) is that for all fitness landscapes  $w$  the induced continuous fitness landscape  $\tilde{w}$  defined over the genotope  $\Pi_{\mathcal{G}}$  is continuous, piecewise linear, and convex, and that the domains of linearity are the cells of a regular polyhedral subdivision  $\Pi_{\mathcal{G}}(w)$  of the genotope. Indeed, for almost all (w.r.t. Borel measure)  $w$ , the subdivisions are even simpler, namely regular triangulations. Thus, the domains of linearity of  $\tilde{w}$  are simplices. The subdivision  $\Pi_{\mathcal{G}}(w)$  induced by  $w$  is called the *shape* of the fitness landscape  $w$ . The set of all polyhedral subdivisions of the genotope (and thus, the set of all shapes of fitness landscapes) is represented by an object called the *secondary polytope*, which can be algorithmically computed. For  $n = 2$ , only two regular subdivisions exist; these correspond exactly to positive and negative epistasis, respectively. For  $n = 3$ , there are 74 shapes of fitness landscapes ([Beerenwinkel et al., 2006](#)).

### 4.1.3 Evolutionary dynamics

The evolutionary model used in this study generalizes the two-locus model of [Christiansen et al. \(1998\)](#) to an  $n$ -locus genotype space  $\mathcal{G} = \{0, 1\}^n$ . We follow the evolution of an infinite (and thus deterministically evolving) population through generations of recombination, mutation, and selection (in this order). These evolutionary events are represented as mappings

$$R, M, S : \Delta_{\mathcal{G}} \longrightarrow \Delta_{\mathcal{G}}$$

that transform a population into a next-generation population according to the rules of the given event.

#### Selection.

In our model, selection is purely fitness-proportionate, so that  $S$  is defined as

$$S(p) = \frac{\langle w, p \rangle}{\|\langle w, p \rangle\|},$$

where the denominator ensures that  $\sum_{g \in \mathcal{G}} p_g = 1$  still holds after selection.

#### Mutation.

Mutation is symmetric, so that a locus may mutate either from 0 to 1 or vice versa. Mutation events at different loci occur independently from each other with constant rate  $\mu$ . Thus, the mutation rate from genotype  $g$  to genotype  $g'$  is given by

$$\mu_{gg'} = \mu^{\sum_i \mathbf{1}_{g_i \neq g'_i}} (1 - \mu)^{\sum_i \mathbf{1}_{g_i = g'_i}}.$$

The effect of mutation on a population can then be written using the linear operator

$$M(p) = Mp = (\mu_{gg'})_{gg'}.$$

#### Recombination.

The possible patterns of recombination given two genotypes  $g^{(1)}$  and  $g^{(2)}$  are defined by the function

$$\begin{aligned} r : \mathcal{G}^2 \times \{1, 2\}^n &\longrightarrow \mathcal{G}, \\ (g^{(1)}, g^{(2)}, k) &\longmapsto g_1^{(k_1)} g_2^{(k_2)} \dots g_n^{(k_n)}. \end{aligned}$$

The index vector  $k$  determines for each locus whether the corresponding locus of the recombinant should carry the genetic information of  $g^{(1)}$  or  $g^{(2)}$ . We always assume  $k_1 = 1$ . Recombination events at different loci occur independently from each other with constant rate  $\rho$ . Thus, the recombinant  $r(g^{(1)}, g^{(2)}, k)$  is produced with rate

$$a_k = \rho^{\sum_{i=1}^{n-1} \mathbf{1}_{k_{i+1} \neq k_i}} \cdot (1 - \rho)^{\sum_{i=1}^{n-1} \mathbf{1}_{k_{i+1} = k_i}}$$

Mating is assumed to be random, so that the fraction of recombinants of genotype  $g$  after recombination is equal to the sum of the individual recombination events between any two genotypes  $g^{(1)}$  and  $g^{(2)}$  which lead to  $g$ , weighted by both the probability of drawing  $g^{(1)}$  and  $g^{(2)}$  from the population and the recombination rate towards  $g$ :

$$R(p) = (\tilde{p}_g)_{g \in \mathcal{G}} \in \Delta_{\mathcal{G}}, \quad \text{with}$$

$$\tilde{p}_g = \sum_{\{g^{(1)}, g^{(2)}, k, l : r(g^{(1)}, g^{(2)}, k) = g\}} p_{g^{(1)}} p_{g^{(2)}} a_k.$$

The combined system.

In discrete time, the sequence of recombination, mutation and selection stages maps a population at generation  $t$  to a population at generation  $t+1$  via the system of difference equations

$$p(t+1) = (S \circ M \circ R)(p(t)).$$

From this time-discrete system, we obtain the corresponding time-continuous version in the standard way (Edelstein-Keshet, 2004): Starting with the discrete identity

$$p(t+1) - p(t) = (S \circ M \circ R - I)(p(t)),$$

we extend the identity to infinitesimally small time steps, and obtain:

$$p'(t) = \lim_{\delta \rightarrow 0} \frac{p(t+\delta) - p(t)}{\delta} = (S \circ M \circ R - I)(p(t)).$$

Example: The system equations in two loci.

In two loci, the mutation matrix is given by

$$M = (\mu_{gg'})_{gg'} = \begin{array}{c} \begin{array}{cccc} & 00 & 01 & 10 & 11 \\ \begin{array}{l} 00 \\ 01 \\ 10 \\ 11 \end{array} & \begin{pmatrix} (1-\mu)^2 & \mu(1-\mu) & \mu(1-\mu) & \mu^2 \\ \mu(1-\mu) & (1-\mu)^2 & \mu^2 & \mu(1-\mu) \\ \mu(1-\mu) & \mu^2 & (1-\mu)^2 & \mu(1-\mu) \\ \mu^2 & \mu(1-\mu) & \mu(1-\mu) & (1-\mu)^2 \end{pmatrix} \end{array} \end{array},$$

and recombination can be written concisely as

$$R(p) = p + \rho \begin{array}{c} \begin{pmatrix} -L \\ L \\ L \\ -L \end{pmatrix} \end{array},$$

with  $L$  being the classical *linkage disequilibrium* (Gillespie, 2004),  $L = p_{00}p_{11} - p_{01}p_{10}$ .

#### 4.1.4 Symmetries of the ODE system

The system of ordinary differential equations introduced above has some interesting invariance properties. To formalize these properties, we write  $S_w$  instead of  $S$  for the selection operator, to emphasize its dependence on the fitness vector  $w$ . We further define, for any permutation  $\sigma \in \text{Sym}(\mathcal{G})$ , the induced mapping

$$\begin{aligned} \chi_\sigma : \mathbb{R}^{\mathcal{G}} &\longrightarrow \mathbb{R}^{\mathcal{G}} \\ (x_{000}, \dots, x_{111}) &\longmapsto (x_{\sigma^{-1}(000)}, \dots, x_{\sigma^{-1}(111)}) \end{aligned}$$

Moreover, let  $\sigma M := \chi_{\sigma^{-1}} \circ M \circ \chi_\sigma$ , and likewise for recombination. For selection, we will also have to permute the fitness vector accordingly:  $\sigma S_w = \chi_{\sigma^{-1}} \circ S_{\chi_\sigma(w)} \circ \chi_\sigma$ . This defines a *group action* (Sagan, 2001) of  $\text{Sym}(\mathcal{G})$  on the set  $\{\sigma f : \sigma \in \text{Sym}(\mathcal{G})\}$ . The symmetries of the ODE system defined by  $f$  are represented by the *stabilizer subgroup*  $\mathcal{S}(f)$  (Sagan, 2001) of  $\text{Sym}(\mathcal{G})$  with respect to  $f$ , i.e. the set  $\{\sigma \in \text{Sym}(\mathcal{G}) : \sigma f = f\}$ . The following proposition establishes these stabilizers for the genetic operators in our system (proofs were performed by exhaustive enumeration in the computer algebra system Mathematica<sup>TM</sup>; the code is available on request).

**Proposition.** *The stabilizer subgroups of  $\text{Sym}(\mathcal{G})$  with respect to selection, mutation, and recombination are given by*

$$\begin{aligned} \mathcal{S}(S_w) &= \text{Sym}(\mathcal{G}), \\ \mathcal{S}(M) &= O_h, \\ \mathcal{S}(R) &= \text{Dih}_4 \times \mathbb{Z}_2. \end{aligned}$$

□

Here,  $O_h$  is the symmetry group of the 3-cube, of order 48. This group consists of 24 rotations, plus one reflection each. It is generated by two rotations and one reflection. The group can be easily visualized by noting that it is isomorphic to  $S_4 \times \mathbb{Z}_2$ . Every permutation within the group is defined by choosing a permutation of the four diagonals in the 3-cube ( $S_4$ ), and by setting a “flag” whether to reflect at a fixed axis or not ( $\mathbb{Z}_2$ ). Thus, informally, permuting the initial conditions of our system will lead to equivalent dynamics if and only if the permutation is an isometry of the 3-cube.

The symmetries of the system under recombination are even more restricted. The group  $\text{Dih}_4$  is the dihedral group of order 8. The product  $\text{Dih}_4 \times \mathbb{Z}_2$  is one of the 14 groups of order 16. Note that  $\text{Dih}_4 \times \mathbb{Z}_2 \subset O_h \subset \text{Sym}(\mathcal{G})$ , such that the symmetries of the full system  $S \circ M \circ R$  are exactly those of the system containing only recombination.

**Example.** We shall explain the previous proposition in the context of a numerical example. Let  $\mu = 0.07$ , and

$$\begin{aligned} w &= (.6739 \ .9994 \ .9616 \ .0589 \ .3603 \ .5485 \ .9618 \ .5973)^t, \\ p(0) &= (.2065 \ .1931 \ .0972 \ .0466 \ .0437 \ .0537 \ .2587 \ .2405)^t. \end{aligned}$$

In addition to the original  $w$  and  $p(0)$  vectors, we consider them in the following permutations:

$$\begin{aligned} \sigma_1 &= (000 \ 100)(001 \ 101)(010 \ 110)(011 \ 111) \in \text{Dih}_4 \times \mathbb{Z}_2, \\ \sigma_2 &= (000 \ 100)(001 \ 110)(010 \ 101)(011 \ 111) \in O_h \setminus \text{Dih}_4 \times \mathbb{Z}_2, \text{ and} \\ \sigma_3 &= (000 \ 100 \ 101 \ 001)(010)(110)(111)(011) \in \text{Sym}(\mathcal{G}) \setminus O_h. \end{aligned}$$

Permutation  $\sigma_1$  is simply a reflection at a parallel of the  $0^{**}$  plane of the 3-cube, whereas  $\sigma_2$  is a reflection at this plane, followed by a reflection at the plane through the vertices 000, 100, 011, and 111. Finally,  $\sigma_3$  rotates the genotypes  $*0^*$ , but leaves all the genotypes  $*1^*$  in place. [Figure 4.1](#) shows the results of simulating the original and permuted systems, for each permutation once without recombination (left column; subfigures a, c, e, g), and once with recombination rate  $R = 1.0$  (right column; subfigures b, d, f, h). As stated in the proposition, without recombination, the dynamics of the system are identical to those of the original system for the permutations  $\sigma_1$  and  $\sigma_2$  (subfigures c, e), because the permutations are from  $O_h$ . With  $\sigma_3 \in \text{Sym } \mathcal{G} \setminus O_h$ , slightly different trajectories and equilibrium states are obtained, as expected. In contrast, with recombination, we only observe identical trajectories for permutation  $\sigma_1 \in \text{Dih}_4 \times \mathbb{Z}_2$  (subfigure f), and the dynamics are already different for  $\sigma_2$ . For  $\sigma_3$  (subfigure h), yet another different set of trajectories is obtained. Note the impact of recombination on the *mutation-selection balance* in permutations  $\sigma_2$  and  $\sigma_3$ , where the most prevalent genotype at equilibrium differs depending on whether  $R = 0$  or  $R = 1$ .

### 4.1.5 Simulations

The simulations described in this section are performed to investigate two questions:

- **The predictive value of shapes on evolutionary outcome (time to equilibrium).** Are some shapes associated with inherently reduced or extended times to system equilibrium?
- **The effect of recombination.** Is there a consistent effect of higher recombination rates on time to equilibrium across all shapes? Is there some consistent effect *within* particular shapes? Or is there no statistically significant association of recombination rate with shapes?

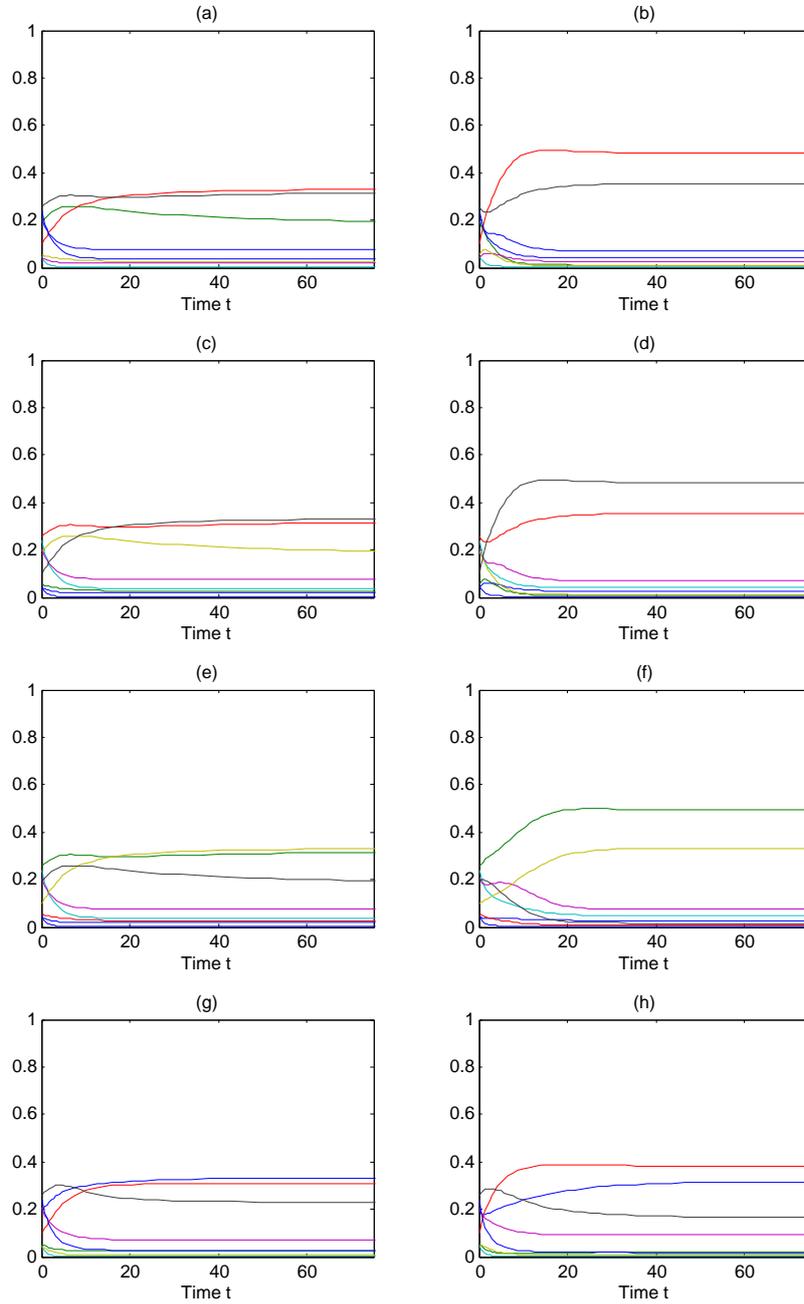


Figure 4.1: The trajectories of the three-locus system defined in the example. The left column shows trajectories without recombination, whereas the right column shows the corresponding system with recombination rate  $R = 1$ . The top row shows the original system, whereas rows 1 – 3 show the effects of applying the permutations  $\sigma_1$ ,  $\sigma_2$ , and  $\sigma_3$  to the fitness vector  $w$ , and to the initial genotype frequencies  $p(0)$ . Different colors indicate different genotypes.

**Simulation setup.** The simulations were performed according to Algorithm 1. A mutation rate  $\mu$  is drawn uniformly from  $[0, 1]$ . An initial population  $p(0)$  is drawn uniformly from the population simplex  $\Delta_{\mathcal{G}} = \{p \in [0, 1]^{\mathcal{G}} : \sum_{g \in \mathcal{G}} p_g = 1\}$ . Uniform distribution on this simplex is represented by the Dirichlet(1,1,1,1,1,1,1,1) distribution.<sup>3</sup> Finally, a fitness vector  $w$  is drawn uniformly from  $[0, 1]^8$ . Using this combination of  $\mu, p(0)$  and  $w$ , the ODE system is then simulated with four different recombination rates 0 (no recombination), 0.1, 0.5, and 1 (maximal recombination). To minimize bias resulting from differences in the relative fitnesses of genotypes in the population, the simulation is replicated in all 40,320 possible permutations from  $\text{Sym}(\mathcal{G})$ . This whole procedure is replicated 1,000 times, leading to a total of 1,000 (replicates)  $\times$  40,320 (permutations)  $\times$  4 (recombination rates) = 161,280,000 simulated systems. The simulations were performed using *Matlab*<sup>TM</sup> with the ode45 solver (code is available on request) on 40 CPUs in parallel for the independent simulation runs. Time to equilibrium was defined as the time point at which all trajectories have entered a 1% margin around their equilibrium.

```

for  $i \leftarrow 1$  to 1000 do
   $\mu \leftarrow$  a random number drawn from the uniform distribution on  $[0, 1]$ ;
   $p(0) \leftarrow$  a random vector drawn from the Dirichlet(1, 1, 1, 1, 1, 1, 1, 1)
  distribution;
   $w \leftarrow$  a random vector from the uniform distribution on  $[0, 1]^8$ ;
  foreach  $\rho \in \{0, 0.1, 0.5, 1.0\}$  do
    foreach  $\sigma \in \text{Sym}(\mathcal{G})$  do
      Simulate the system using  $\mu, \rho, \sigma w, \sigma p(0)$ ;
      Record for this set of parameters the time until the system is in
      equilibrium;
    end
  end
end

```

**Algorithm 1:** Simulating trajectories for random populations.

**Results.** Figure 4.2 shows the results of our simulation study. For each of the 74 shapes of fitness landscapes, the mean time to equilibrium for the simulated populations is indicated by the height of the corresponding bar. Black bars correspond to recombination rates of 0, whereas increasingly lighter gray tones indicate recombination rates of 0.1, 0.5, and 1.0, respectively. Each of the 161,280,000 population

<sup>3</sup>The naive approach of sampling eight numbers uniformly from  $[0, 1]$ , and then normalizing them to one would not yield the desired uniform distribution on the population simplex.

simulations contribute to exactly one of the bars. The results can be summarized as follows:

- For each of the four recombination rates  $\rho \in \{0, 0.1, 0.5, 1.0\}$ , the mean time to equilibrium strongly depends on the shape of the fitness landscape. For example, in the absence of recombination, the mean time to equilibrium in shape 74 is only little more than ten units of time. In contrast, a mean of 60 time units is necessary in shapes 20–25 to reach equilibrium in the absence of recombination. The visual impression is confirmed by the Kruskal-Wallis test, which tests for the null hypothesis that the time to equilibrium is the same in the 74 different shapes. For all four recombination rates, this null hypothesis can be rejected with extremely high significance ( $p < 10^{-16}$ ).
- Within all shapes, presence of recombination leads to shorter times to equilibrium in mean. However, the relation between recombination rate and mean time to equilibrium appears to be non-monotonic, the minimum being reached at intermediate recombination rates.
- The speed-up in time to equilibrium mediated by recombination varies across the different shapes of fitness landscapes.

#### 4.1.6 Comparison with a rank ordering classification of fitness landscapes

Another classification system for fitness landscapes was recently proposed by [Weinreich \(2005\)](#). Briefly, the classes of fitness landscapes in that system are based on the rank ordering of the genotypes according to their fitness values. For example, all fitness vectors  $w$ , for which  $w_{000} < w_{001} < \dots < w_{111}$  would be grouped into the same class according to their common rank ordering of genotypes. Symmetry considerations reduce the number of distinct classes to below the theoretically possible number of  $|\mathcal{G}|!$ . In particular, [Weinreich \(2005\)](#) focuses on fitness landscapes lacking sign epistasis, giving rise to eight different classes of landscapes for three loci. In this section, we compare this combinatorial classification to the geometric classification based on the shape of the landscape proposed by [Beerenwinkel et al. \(2006\)](#).

Fitness landscapes were generated by drawing eight samples from the uniform distribution on  $[0, 1]$ . These were then permuted such that the eight classes of [Weinreich \(2005\)](#) are sampled uniformly. [Table 4.1](#) shows both classifications for 10,000 fitness landscapes generated according to this scheme. Interestingly, even though the sampling scheme was restricted to landscapes lacking sign epistasis, all of our 74 shapes are hit by samples. Nevertheless, a small number of shapes receive the largest number of hits.

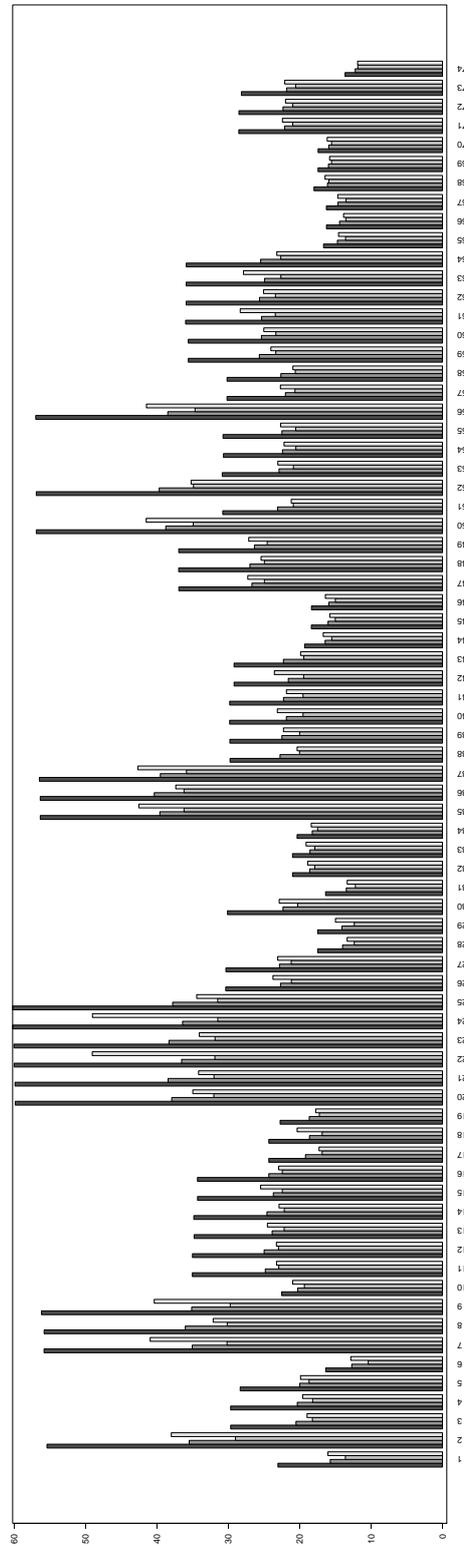


Figure 4.2: Time until reaching equilibrium within the different shapes of fitness landscapes, for different recombination rates  $\rho$ . The rates  $\rho \in \{0, 0.1, 0.5, 1.0\}$  are indicated by increasingly lighter gray scale.

Table 4.1: Sampling 10,000 fitness landscapes lacking sign epistasis, as described in the text. Rows indicate the eight classes of fitness landscapes according to Weinreich, while columns indicate the 74 shapes of fitness landscapes according to the classification of [Beerenwinkel et al. \(2006\)](#). Entry  $(i, j)$  in the table indicates the number of fitness landscapes of Weinreich class  $i$  and shape  $j$ .

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	21	10	4	28	0	6	9	3	0	0	54	136	0	15	0	0	0	0	0	54
2	54	13	11	8	17	1	21	5	1	0	75	49	6	9	13	20	1	1	4	23
3	2	105	2	0	0	0	131	0	0	0	13	0	1	0	0	0	0	0	0	162
4	9	68	1	7	0	1	20	28	0	3	13	41	0	2	0	0	0	1	0	104
5	11	114	2	3	5	2	52	20	6	0	19	19	2	4	3	11	1	1	0	70
6	2	24	5	0	0	1	33	0	0	2	78	0	0	0	0	0	3	0	0	155
7	11	47	4	0	1	1	82	0	3	1	24	0	18	0	15	0	2	0	0	64
8	3	36	0	0	0	0	23	7	0	0	50	78	0	0	0	0	0	1	0	194
	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	19	0	3	0	0	0	4	0	6	0	0	0	0	0	135	89	0	11	21	0
2	10	3	6	1	3	3	2	0	0	2	4	4	0	2	220	117	20	5	3	0
3	0	22	0	0	0	9	0	1	0	0	0	4	0	0	454	0	0	2	0	0
4	76	0	13	0	0	2	13	0	3	0	0	0	8	0	177	205	0	4	11	0
5	43	11	11	2	8	5	0	0	1	4	0	2	1	4	298	146	41	5	0	0
6	0	0	0	0	0	13	0	1	0	0	0	3	0	0	179	0	0	26	0	0
7	0	58	0	9	0	11	0	1	0	2	1	2	0	0	371	0	23	3	0	2
8	59	0	0	0	0	4	5	0	0	0	0	0	0	0	208	70	0	24	40	0
	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
1	19	0	3	0	0	0	4	0	6	0	0	0	0	0	135	89	0	11	21	0
2	10	3	6	1	3	3	2	0	0	2	4	4	0	2	220	117	20	5	3	0
3	0	22	0	0	0	9	0	1	0	0	0	4	0	0	454	0	0	2	0	0
4	76	0	13	0	0	2	13	0	3	0	0	0	8	0	177	205	0	4	11	0
5	43	11	11	2	8	5	0	0	1	4	0	2	1	4	298	146	41	5	0	0
6	0	0	0	0	0	13	0	1	0	0	0	3	0	0	179	0	0	26	0	0
7	0	58	0	9	0	11	0	1	0	2	1	2	0	0	371	0	23	3	0	2
8	59	0	0	0	0	4	5	0	0	0	0	0	0	0	208	70	0	24	40	0
	61	62	63	64	65	66	67	68	69	70	71	72	73	74						
1	0	5	0	0	3	0	2	1	0	4	45	90	0	1						
2	1	8	0	9	0	0	1	0	1	6	96	16	7	2						
3	4	0	0	0	0	0	0	1	1	0	66	0	0	0						
4	0	8	0	0	1	0	0	0	0	1	19	42	0	2						
5	1	2	0	2	1	0	0	0	0	4	29	4	3	4						
6	0	0	0	0	1	0	0	4	0	0	186	0	0	1						
7	23	0	8	0	2	3	0	1	2	0	95	0	6	4						
8	0	0	0	0	1	0	0	4	0	0	43	20	0	2						

Visual inspection of [Table 4.1](#) suggests a low degree of similarity among the two classification systems. However, it is desirable to provide a more quantitative characterization of similarity. The standard way of comparing partitions (such as those induced by the two classifications of fitness landscapes) is based on the *adjusted Rand index* ([Hubert and Arabie, 1985](#)). The Rand index of the partition shown in [Table 4.1](#) is 0.04. With 1 indicating maximal agreement, and 0 completely random agreement, this confirms the visual impression of two virtually unrelated classifications.

### 4.1.7 Discussion

The analytical intractability of all but the simplest population genetic models involving mutation, selection, and recombination has stimulated the search for features that are to some degree associated with the fate of a population. Intuitively, since the fitness landscape governs the selection process, it should have a substantial impact on the fate of a population, along with initial conditions and mutation or recombination rates. This intuition has been confirmed in various studies when focusing on pairs of loci using the concept of epistasis. In this study, we have investigated whether a novel extension of classical epistasis to multilocus systems proposed by [Beerenwinkel et al. \(2006\)](#) exhibits a similarly pronounced association with the fate of a population. We have focused on a deterministically evolving biallelic three-locus system, and the property of “time to equilibrium“ as a proof of concept.

We have proven that the 74 classes of fitness landscapes existing in this case are related to each other via symmetries determined by the definition of the genetic operators. Moreover, using large-scale numerical simulations, we have shown that the shape of the fitness landscape is significantly associated with the time to equilibrium at any of the tested recombination rates. The presence of recombination has led to a shorter mean time to equilibrium within all shapes, although the magnitude of the effect appears to be shape-dependent. The consistent benefit of recombination (albeit of varying size) is surprising in the light of theories for the two-locus case that suggest a benefit of recombination mostly in the case of negative epistasis. Recombination might have been expected to have beneficial effects in some shapes and detrimental effects in others, in analogy to the case of negative and positive epistasis, respectively. However, these theories are linked to very specific population genetic models (often making use of modifier loci influencing recombination) and do not generalize well to a wide range of scenarios ([Otto and Lenormand, 2002](#); [Kouyos et al., 2006](#)). Finally, we have compared the shape-based classification of fitness landscapes by [Beerenwinkel et al. \(2006\)](#) to the rank ordering-based classification by [Weinreich \(2005\)](#). Our results show that the two classification systems are virtually orthogonal to each other, meaning that they focus on very different features of the fitness landscape. Combined with the observation that both classifications are strongly associated with evolutionary dynamics, it seems desirable to investigate the use of a hybrid classification combining the two

individual systems. Due to their orthogonality, this hybrid classification scheme would likely be a stronger predictor of evolutionary dynamics than each individual scheme.

## 4.2 Mutation dynamics during interruption of antiretroviral therapy<sup>4</sup>

In this section, we study the dynamics of mutation disappearance in the absence of drug pressure, based on longitudinal data collected during clinical practice. While treatment interruptions (TI) are not a recommended therapeutic strategy, unstructured TIs are part of the every-day life of many patients (e.g. due to non-adherence or insufficient drug levels). Our analysis strongly suggests that mutation loss does not revert the complex patterns of mutation accumulation observed during antiretroviral therapy. Rather, mutations are shown to disappear largely independently from each other, albeit at individual rates.

### 4.2.1 Introduction

Combination therapy against human immunodeficiency virus HIV-1 can substantially delay disease progression, prolong survival, and maintain quality of life. However, treatment cannot clear the virus from the patient and, if viral load is to remain suppressed, drug administration needs to occur indefinitely. Moreover, the duration of treatment response is limited and highly variable. Therapy failure is primarily caused by the emergence of pre-existing or newly produced drug-resistant viral variants. The patterns of resistance mutations accumulating over time in the presence of drugs have been studied extensively (cf. Section 5.1). In general, mutations become fixated within the viral quasi-species in a non-deterministic order. However, positive or negative interactions between mutations (cf. Section 5.2) lead to stochastic preferences for evolution to proceed along more or less well-defined resistance pathways (cf. Section 4.3).

In response to treatment with nucleoside reverse transcriptase inhibitors (NRTIs), the most prominent positive interactions are known to occur among the nucleoside analogue mutations (NAMs). These mutations enhance the phosphorolytic removal of the chain-terminating NRTI from the 3'-terminus of the primer ("primer unblocking"). The classical NAMs fall into two distinct classes: NAM I (consisting of the core mutations M41L, L210W, and L215Y) and NAM II (D67N, K70R, K219Q). While a substantial overlap exists between these classes, several studies have confirmed that mutations from the same NAM class co-occur significantly more frequently than expected under an independence assumption, and mutations from different NAM classes significantly less frequently (Svicher et al., 2006; Sing et al., 2005b; Cozzi-Lepri et al.,

---

<sup>4</sup>The work reported in this section was performed in collaboration with Valentina Svicher, Francesca Ceccherini-Silberstein, and Carlo-Federico Perno, University of Rome "Tor Vergata".

2005; Gonzales et al., 2003). A strong antagonism exists between the NAM mutations and point mutations like K65R or M184V. The latter mutations are involved in a different resistance mechanism which leads to better discrimination between NRTIs and the corresponding dNTP substrate. Another cluster, less frequently observed, but equally well defined as the NAMs, consists of mutations at the RT positions 151, 62, 75, 77, and 116.

With a wealth of work on the response of a wild-type population to drug-induced change in the natural fitness landscape of HIV, it is remarkable how little is known about the mutational dynamics of a resistant viral population during interruption of drug pressure. Clinically, structured treatment interruptions (STI) have been expected to boost host cellular response, minimize drug toxicities, and, perhaps most importantly, allow the viral population to revert back to wild-type. It was expected that such a reversion would improve the range of future drug options or the virological response to follow-up therapy (reviewed in Benson, 2006). Unfortunately, according to a number of studies, STI in patients with multidrug-resistant HIV was found to have no clinical or virological benefit and thus should not be applied (Benson, 2006). Still, some findings indicate benefits of interrupting an individual drug in a class for which there has been an accumulation of resistance mutations. However, the existence of latent reservoirs in which virtually every resistant variant becomes archived presents a fundamental limitation to approaches that aim to revert resistance mutations in the viral population.

Despite these limitations, the study of mutational dynamics of resistant populations in the absence of drug pressure is crucially interesting from at least two points of view. From the clinical point of view, we have to accept the fact that even if STI might not be clinically advised, "unstructured" therapy interruptions occur every day as part of life for many HIV-infected patients, for a wide variety of reasons (e.g. non-adherence or toxicities). From the scientific point of view, the study of therapy interruptions is particularly interesting since it represents the dual of the standard situation of drug therapy: In the latter case, a naive population is subjected to a fitness landscape shaped primarily by drug factors. In the former case, a resistant population is subjected to a fitness landscape shaped primarily by host factors. Will the resulting evolutionary pathways also be duals of each other?

In this study, we focus specifically on the dynamics of classical NRTI mutations, plus a number of "novel" NRTI-associated mutations (Svicher et al., 2006), during treatment interruption. To our knowledge, this study is the first application of methods from survival analysis to longitudinal sequence data. First, we compare survival functions for individual mutations. Next, we analyze if the survival function of a mutation depends on the presence of specific other mutations at baseline. Based on these analyses, we propose a model for mutation dynamics during TI and compare it to a naive linear model.

## 4.2.2 Materials and Methods

### Patient population

The study included 132 HIV-1 infected, drug-experienced adult patients treated in clinical centers in or around Rome, and undergoing genotypic resistance testing for routine clinical purposes between 1999 and 2006. Data for all patients were stored in a specifically designed anonymous database that included mutational, demographic, immunologic, virologic and therapeutic parameters. The decision for TI was either based on limited remaining treatment options or on patient choice (mainly due to side effects of the drugs). The median time of exposure to antiretroviral treatment before TI was 7.3 years. At the time of starting interruption, the patients had been exposed to an average of 5 (IQR=4–5) NRTIs, 2 (IQR=1–4) protease inhibitors (PIs), and 1 (IQR=1–1) (non-nucleoside RT inhibitor) NNRTIs and their median number of resistance mutations were: 4 (IQR=3–5) for NRTIs, 10 (IQR=8–10) for PIs and 2 (IQR=1–2) for NNRTIs. A total of 144 treatment interruption episodes (TIEs) were available for the 132 patients (nine patients underwent two, and one patient four TIEs). For each TIE, a baseline sample was taken before TI onset (days before onset: mean [median]=33.21 [31.00]; IQR: 16.00–47.50; min–max: 0–90). A baseline genotype was obtained in all cases as described below, and baseline viral load (VL) and CD4<sup>+</sup> cell counts were available in 134 and 130 cases, respectively.

### HIV sequencing

HIV genotype analysis was performed on plasma samples by means of a commercially available kit (the ViroSeq™ HIV-1 Genotyping System, versions 1 and 2, Applied Biosystems) and an automatic sequencer (ABI 377 and ABI 3100, Applied Biosystems, Foster City, California, USA) (Perno et al., 2001; Ceccherini-Silberstein et al., 2004). Briefly, RNA was extracted, retrotranscribed by MULV RT, and amplified with Amplitaq-Gold polymerase enzyme by using two different sequence-specific primers for 40 cycles. Pol-amplified products (containing the entire protease and the first 335 amino acids of the reverse transcriptase open reading frame) were full-length sequenced in sense and antisense orientations by using seven different overlapping sequence-specific primers for an automated sequencer (ABI 3100). Mixtures of multiple residues at a single position were all considered as present in the statistical analyses. The isolates were subtyped by comparing them to reference sequences of known subtype (<http://hivdb.stanford.edu>). All sequences were found to be of subtype B.

### Mutations

The mutations considered in the analysis were those reported to be associated with resistance to nucleoside reverse transcriptase-inhibitors (NRTIs) in the mutation lists

of the International AIDS Society (Johnson et al., 2006), the Stanford HIV Drug Resistance Database<sup>5</sup>, and in a comprehensive screening for “novel” mutations (Svicher et al., 2006). Specifically, we focused on 35 NRTI resistance mutations, 20 classical (M41L, E44D, A62V, K65R, D67N, K70R, L74V, V75I, F77L, Y115F, F116Y, V118I, Q151M, M184I, M184V, L210W, T215F/Y, K219E/Q), and 15 novel (K20R, V35M, T39A, K43E/N/Q, I50V, R83K, K122E, G196E, E203D/K, H208Y, F214L, D218E). We will denote the set of these 35 mutations by the symbol  $\mathcal{M}$ . Of the mutations considered, a mean (median) number of 6.6 (6.0) was present at baseline (IQR=4.0–8.3; min–max=1.0–15.0). At follow-up, the mean (median) number of mutations was 4.1 (2.0), with IQR=1.0–5.3 and min–max=0.0–21.0.

### Survival analysis for mutation dynamics

For all mutations  $m \in \mathcal{M}$ , survival functions  $S_m$  were estimated using the Kaplan-Meier product-limit estimator. For a given mutation  $m$ , all therapy interruptions with  $m$  present at baseline were considered. Observations were treated as right-censored, i.e. if mutation  $m$  was present in a sequence from time  $t_1$ , and absent in a sequence from time  $t_2 > t_1$ , mutation  $m$  was assumed to disappear exactly at time  $t_2$  (cf. Discussion). If a mutation present at baseline was still present in the last follow-up within a TI period at time  $t$ , a survival time of “at least  $t$ ” was assumed.

The impact of baseline viral load, CD4<sup>+</sup> cell count, and baseline number of mutations on the survival function of a mutation was assessed using Cox proportional hazards regression. To assess if the survival function of a mutation  $m$  depends on the baseline presence of some other mutation, we proceeded as follows: For each mutation  $m' \in \mathcal{M} \setminus \{m\}$ , the TI episodes with  $m$  present at baseline were split into two classes, according to whether  $m'$  was present or absent at baseline. Two separate survival curves for  $m$  were then estimated from the two classes. The null hypothesis that the two survival curves are identical was tested using the log-rank test. This procedure was only performed if at least five TI periods were available in each of the two classes. Correction of p-values for multiple testing was performed using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

### Independence model of mutation loss

As will be shown below, the results from the previous analyses suggest that mutations are lost independently from each other. For confirmation, we evaluated the fit of an independence model that predicts the number of mutations present after a given time post TI onset from the set of mutations present at baseline. The model is based on the individual survival functions  $S_m$  for the mutations  $m \in \mathcal{M}$ . For each time  $t > 0$  after onset of therapy interruption,  $S_m(t)$  gives the probability that mutation  $m$  is still

<sup>5</sup><http://hivdb.stanford.edu>

present at time  $t$ , provided it was present at baseline. Let  $(\mathbf{M}_t)_{t \geq 0}$  be a family of random variables denoting the mutations present at time  $t$ . Then, assuming independent mutation disappearance,

$$P(\mathbf{M}_t = M : \mathbf{M}_0 = M_0) = \prod_{m \in M} S_m(t) \cdot \prod_{m \in M_0 \setminus M} (1 - S_m(t)),$$

for any  $m_0 \subset m \subset \mathcal{M}$ . It follows that

$$P(|\mathbf{M}_t| = k : \mathbf{M}_0 = M_0) = \sum_{M \subset M_0: |M|=k} P(\mathbf{M}_t = M : \mathbf{M}_0 = M_0).$$

Using a linearity argument familiar from e.g. the expectation of the binomial distribution, the expected number of mutations at time  $t$  given the set of mutations  $M_0$  at baseline is simply given as

$$E[|\mathbf{M}_t| : \mathbf{M}_0 = M_0] = \sum_{m \in M_0} S_m(t).$$

### Discordance between predicted and observed mutation counts

To identify situations in which the disagreement between predicted and actually observed mutation counts is largest, we fitted linear regression models for the dependent variable "expected number of mutations - observed number of mutations". In simple linear regression, we evaluated the following potential predictors of discordance: baseline viral load; CD4+ cell count; time  $t$  after onset of TI; number of mutations that had actually disappeared until time  $t$ . The significance of these predictors in combination was assessed using multiple linear regression.

### Linear model of mutation loss

The independence model of mutation loss is quite simple as it assumes no interactions between mutations. However, it allows each mutation to have an individual, and possibly non-linear survival function. To compare whether this added flexibility is necessary and adequate, we also fitted a linear model predicting the number of mutations at time  $t$  only based on the number of mutations at baseline.

## 4.2.3 Results

### Baseline patient characteristics

The study population was predominantly male (103/138 patients, 74.6%), and had a median age of 40 years at baseline (IQR=36–45). At baseline, mean (median) CD4<sup>+</sup> count was 514.0 (489.0) cells per mm<sup>3</sup> (IQR=327.8–642.8; min–max=21.0–1600.0). Mean (median) VL at baseline was 40,930 (7552) copies per ml (IQR=2227–28590; min–max=105–810000).

### Follow-up data during therapy interruption

For 114 of the 144 patients, a single follow-up during treatment interruption was available. Two follow-ups were available for 25 patients. Three and four follow-ups were given for four and one patients, respectively. Mean (median) follow-up time was 5.4 (4.3) months, with an IQR from 3.1 to 6.8 months, a minimum of one month and a maximum of 28.8 months.

### Emergence of NRTI resistance mutations not present at baseline

Some studies have reported anecdotal evidence on the appearance of resistance mutations not present/detected at baseline during the TI. In our data, this seemingly counter-intuitive (but see Discussion section) phenomenon of emerging mutations not detected at baseline was very rare for resistance-associated mutations: 44D (1 TBE), 70R (3), 118I (3), 184V (1), 210W (1), 215Y (2), 20R (1), 35M (2), 39A (2), 43E (2), 50V (5), 83K (15), 122E (3), 196E (1), 203D (1), 208Y (2), 214L (7). Thus, follow-up samples from within the TIE represent almost exclusively maintenance or loss of mutations already present at baseline.

### Survival analysis for individual mutations

Table 4.2 shows statistics for the estimated survival functions of the 35 mutations considered here. The median survival times for all classical NRTI resistance mutations lie between 3.1 (184V) and 6.8 (116Y) months. The survival time of novel mutations associated with NRTI failure also lies within this range. In contrast, resistance-associated polymorphisms (122E, 196E, 214L) show a substantially longer median survival (8.0–11.6). Finally, the mutations 50V and 83K which have been shown to be negatively associated with NRTI failure (Svicher et al., 2006) do not disappear at all in the absence of drug pressure. Finite 95% upper confidence intervals could only be calculated for sufficiently frequent mutations.

Using Cox proportional hazards regression, we then analyzed the effect of baseline VL, CD4<sup>+</sup> cell count, and baseline number of mutations (each in an individual Cox model) on the survival curve of each individual mutation. Baseline number of mutations was never a significant predictor. Similarly, baseline CD4<sup>+</sup> cell count was significant only for 122E ( $p=0.037$ ), and with a very weak effect size. In contrast, baseline viral load was significantly associated with mutation survival for several mutations: 196E ( $p = 0.017$ ), 20R ( $p = 0.006$ ), 215Y ( $p = 0.050$ ), 184V ( $p = 2.4 \cdot 10^{-6}$ ), 118I ( $p = 0.001$ ), and 44D ( $p = 0.040$ ). The estimated effect size was relatively consistent across the mutations in that an increase in VL by 10,000 copies decreased the survival probability of a mutation by about 10%. The lack of significance for other mutations may possibly be due to a lack of data.

Table 4.2: Statistics for the estimated survival functions of the 35 mutations considered.

Mutation	N (Events)	Mean (SD)	Median (95% CI)
41L	68 (46)	5.61 (0.37)	4.70 (3.90 – 6.30)
44D	25 (15)	5.22 (0.48)	4.80 (3.60 – $\infty$ )
62V	10 (6)	7.63 (1.64)	5.60 (5.20 – $\infty$ )
65R	8 (8)	6.11 (1.01)	5.70 (4.00 – $\infty$ )
67N	84 (63)	5.25 (0.34)	4.40 (3.80 – 5.60)
70R	59 (43)	5.68 (0.43)	5.20 (4.40 – 6.80)
74V	32 (27)	5.55 (0.53)	4.40 (3.90 – 7.40)
75I	8 (7)	6.14 (1.44)	4.90 (4.20 – $\infty$ )
115F	8 (8)	6.03 (1.20)	5.20 (4.00 – $\infty$ )
116Y	7 (7)	7.89 (1.25)	6.80 (5.50 – $\infty$ )
118I	35 (22)	5.97 (0.56)	4.90 (4.20 – 8.00)
151M	9 (8)	6.55 (1.13)	5.60 (4.20 – $\infty$ )
184I	5 (5)	3.44 (0.45)	3.10 (2.70 – $\infty$ )
184V	95 (87)	5.09 (0.31)	4.30 (3.90 – 5.00)
210W	45 (27)	6.33 (0.49)	6.30 (4.10 – 8.00)
215F	36 (28)	4.76 (0.43)	4.20 (3.50 – 5.60)
215Y	61 (44)	5.69 (0.35)	5.20 (4.10 – 7.30)
219E	14 (10)	5.91 (0.76)	5.20 (3.90 – $\infty$ )
219Q	41 (29)	5.58 (0.55)	5.20 (4.30 – 6.50)
20R	26 (13)	5.75 (0.54)	5.60 (3.80 – $\infty$ )
35M	15 (6)	14.19 (3.87)	6.30 (5.20 – $\infty$ )
39A	17 (9)	6.11 (0.82)	4.60 (3.30 – $\infty$ )
43E	9 (5)	4.16 (0.70)	3.60 (3.30 – $\infty$ )
43N	5 (2)	8.88 (1.67)	10.80 (3.10 – $\infty$ )
43Q	5 (3)	5.85 (1.14)	8.00 (3.90 – $\infty$ )
50V	5 (0)	8.40 (0.00)	$\infty$ ( $\infty$ – $\infty$ )
83K	28 (1)	22.75 (0.74)	$\infty$ ( $\infty$ – $\infty$ )
122E	60 (19)	8.65 (0.79)	8.30 (6.50 – $\infty$ )
196E	50 (9)	10.76 (0.92)	11.60 (9.40 – $\infty$ )
203K	14 (9)	5.36 (0.97)	5.10 (3.10 – $\infty$ )
208Y	13 (7)	5.91 (0.77)	5.70 (4.40 – $\infty$ )
214L	30 (11)	8.36 (0.95)	8.00 (5.20 – $\infty$ )
218E	19 (14)	4.75 (0.44)	4.50 (3.50 – 7.40)

Table 4.3: The seven mutation pairs with significant ( $p < 0.05$ , log-rank test) interaction in mutation survival. For example, row one compares the survival function of 122E, depending on whether 219Q is not present (stratum 0) or present (stratum 1). The median survival time in the absence of 219Q is 12.6 months, while the median in the presence of 219Q decreases to 6.5 months. The estimated survival function of this most significant interaction is shown in [Figure 4.3](#). None of the interactions remained significant after correcting for multiple testing.

Mutation	Stratification	stratum size (0;1)	Median survival in stratum (0;1)	p (log-rank)
122E	219Q	35; 25	12.6; 6.5	0.002
196E	44D	36; 14	11.6; $\infty$	0.049
70R	67N	7; 52	6.8; 4.6	0.046
184V	67N	39; 56	4.5; 4.0	0.043
122E	67N	17; 43	12.6; 7.4	0.032
122E	70R	33; 27	9.7; 6.5	0.035
214L	219Q	13; 17	9.7; 5.2	0.040

### Interaction effects in mutation survival

For each pair of mutations  $m \neq m' \in \mathcal{M}$ , we estimated the survival function of  $m$  depending on whether  $m'$  is present at baseline or not. This procedure was only performed if from all samples with  $m$  at baseline, at least five samples were available with  $m'$  present, and five with  $m'$  absent. This was the case for 135 of the theoretically possible 1,190 pairs. Before correction for multiple testing, the p-values appear very close to uniformly distributed in the interval  $[0, 1]$  (mean [median] 0.50 [0.48], IQR=0.25–0.67, min–max=0.00–1.00). In a multiple testing problem, this usually indicates that in none of the cases the null hypotheses can be rejected. Indeed, after correcting p-values for multiple testing using the Benjamini-Hochberg method, all of the potential interactions became insignificant ( $p > 0.29$ ). In [Table 4.3](#), we show the mutation pairs with a p-value smaller than 0.05 before correction for multiple testing.

Of course, the inability to reject the null hypothesis that survival curves of a mutation differ depending on the presence or absence of another mutation does not constitute “proof” of independence. For example, in [Figure 4.3](#), we have contrasted the survival times of 122E and 184V, in both cases stratified by the baseline presence of 219Q. While the visual impression for 184V confirms the lack of interaction  $p = 0.944$  with 219Q, the 219Q-stratified survival curves for 122E are remarkably different, despite the lack of significance ( $p = 0.29$ ) after correcting for multiple testing.

Additionally, in theory, the absence of significant pair-wise interactions does not preclude the possible existence of higher-order interactions (i.e. between triplets or even larger groups of mutations). In practice, however, this is highly implausible.

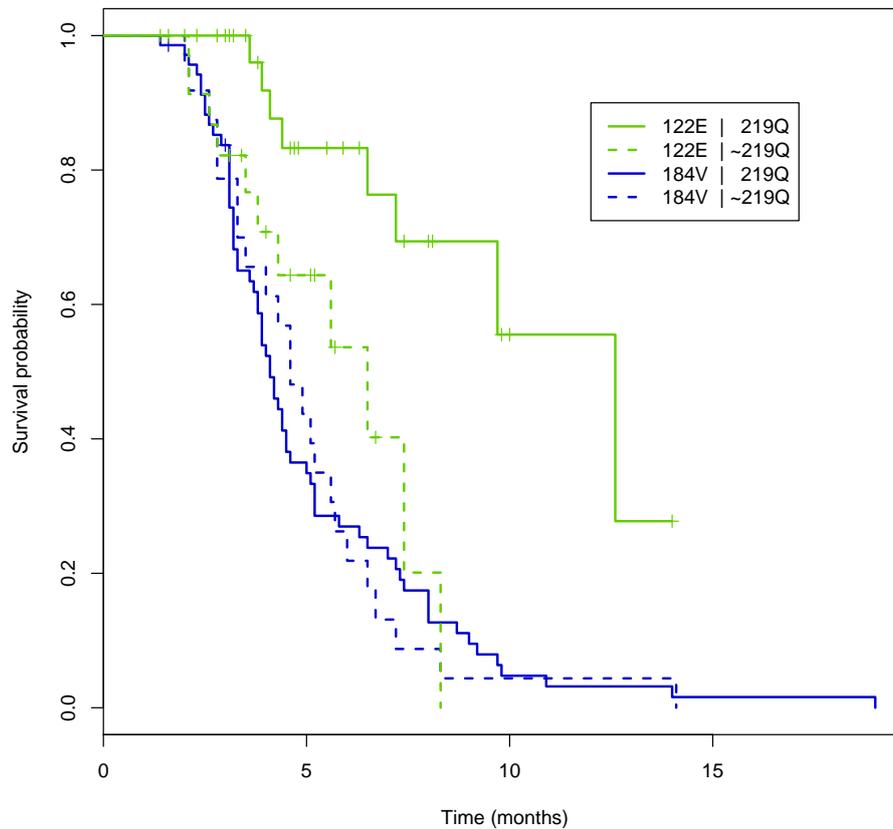


Figure 4.3: Estimated survival functions for M184V (blue) and K122E (green), depending on whether K219Q is present at baseline (solid) or not (dashed). The K122E/K219Q interaction was the most significant among all pairs of mutations ( $p = 0.0021$ ), while M184V/K219Q is shown as an example of a clearly non-significant interaction ( $p = 0.944$ ). As mentioned in the text, none of the interactions remained significant after correcting for multiple testing. Censoring times are marked by a vertical dash.

Moreover, there will rarely be enough longitudinal data to exhaustively test for dependencies of order three or more. Given the infeasibility of exhaustive higher-order testing, we instead tested special cases in which higher-order interactions might be expected. Specifically, we compared the survival probability of the core TAM mutations in the presence or absence of mutations from the other TAM group, and in the presence or absence of mutations from the extended TAM cluster. As shown in [Table 4.4](#), there was a very weak trend that in both variants the additional mutations did not extend, but rather limited the survival time of the mutations. However, none of the observed differences turned out to be significant.

### Goodness of fit of the independence model

The results from the previous section strongly suggest that if pair-wise interactions play any role at all in the speed of mutation disappearance, that this role is very marginal. Higher-order interactions are infeasible to be tested exhaustively, but the manually selected test scenarios shown in [Table 4.4](#) show nothing significant either. Thus, disregarding all non-sequence information (host factors, clinical parameters), mutation dynamics in viral load should be predictable using an independence model (cf. Methods). By considering the presence or absence of a mutation  $m$  at a specific time  $t$  as the outcome of a Bernoulli experiment with success probability given by the value of the estimated survival function of  $m$  at time  $t$ , our goal is to predict, given a set of mutations present at baseline, the number of remaining mutations at a given time after onset of the therapy break.

The absolute difference between observed and expected (according to the independence model) number of mutations was 1.23 in median (IQR 0.39-2.74). In contrast, the absolute difference between baseline number of mutations and observed number of mutations (i.e. a null model assuming no mutation dynamics at all) was 3.00 in median (IQR 1.00-6.00). To obtain a second comparison, we fitted a linear model to the data for predicting the number of mutations at time  $t$  from the number of mutations at baseline and the time  $t$ . Ignoring all mutation-specific information (in contrast to the independence model which relies on mutation-specific survival probabilities), this model led to a median difference of 1.84 mutations (IQR 0.74-2.93), despite the advantage given to it by being fitted directly to the data for which predictions were made.

### Predictors of discordance between independence model and observed number of mutations

In an attempt to characterize situations of discordance between observed mutation counts and those expected under the independence model, we fitted linear regression models for the target variable "expected number of mutations - true number of muta-

Table 4.4: Analysis of potential higher-order interactions between mutations in mutation survival. The first column shows the mutations required at baseline for a sample to be included in an analysis. The second column denotes the endpoint used in survival analysis. For example, in the first row the endpoint is the loss of the first of the three mutations considered). Column three contains the stratification criterion, column four the number of samples in each stratum, column five and six the (restricted) mean and median survival time for each stratum, and column seven contains the p-value from the log-rank test.

Baseline	Endpoint	Stratification	N	Mean	Median	p
41L, 210W, 215Y	41L + 210W + 215Y $\leq$ 2	67N + 70R + 219Q $\geq$ 0	15; 27	5.95; 5.56	6.5; 4.3	0.91
41L, 210W, 215Y	41L + 210W + 215Y $\leq$ 1	67N + 70R + 219Q $\geq$ 0	15; 27	5.95; 6.23	6.5; 4.7	0.76
41L, 210W, 215Y	41L + 210W + 215Y = 0	67N + 70R + 219Q $\geq$ 0	15; 27	5.95; 6.23	6.5; 4.7	0.98
41L, 210W, 215Y	41L + 210W + 215Y $\leq$ 2	39A+ 43E+ 43Q+ 122E+ 203K+ 208Y $\geq$ 0	6; 36	7.15; 5.27	7.1; 4.7	0.17
41L, 210W, 215Y	41L + 210W + 215Y $\leq$ 1	39A+ 43E+ 43Q+ 122E+ 203K+ 208Y $\geq$ 0	6; 36	7.15; 5.52	7.1; 4.8	0.35
41L, 210W, 215Y	41L + 210W + 215Y = 0	39A+ 43E+ 43Q+ 122E+ 203K+ 208Y $\geq$ 0	6; 36	7.15; 5.55	7.1; 4.8	0.44
67N, 70R, 219Q	67N + 70R + 219Q $\leq$ 2	41L + 210W + 215Y $\geq$ 0	26; 7	4.58; 3.89	4.3; 3.8	0.28
67N, 70R, 219Q	67N + 70R + 219Q $\leq$ 1	41L + 210W + 215Y $\geq$ 0	26; 7	4.88; 3.89	4.6; 3.8	0.15
67N, 70R, 219Q	67N + 70R + 219Q = 0	41L + 210W + 215Y $\geq$ 0	26; 7	5.54; 4.2	6.1; 3.8	0.14
67N, 70R, 219Q	67N + 70R + 219Q $\leq$ 2	218E+ 214L $\geq$ 0	10; 23	4.89; 4.21	4.2; 4.3	0.30
67N, 70R, 219Q	67N + 70R + 219Q $\leq$ 1	218E+ 214L $\geq$ 0	10; 23	4.89; 4.51	4.2; 4.6	0.52
67N, 70R, 219Q	67N + 70R + 219Q = 0	218E+ 214L $\geq$ 0	10; 23	5.53; 5.2	6; 5.7	0.78

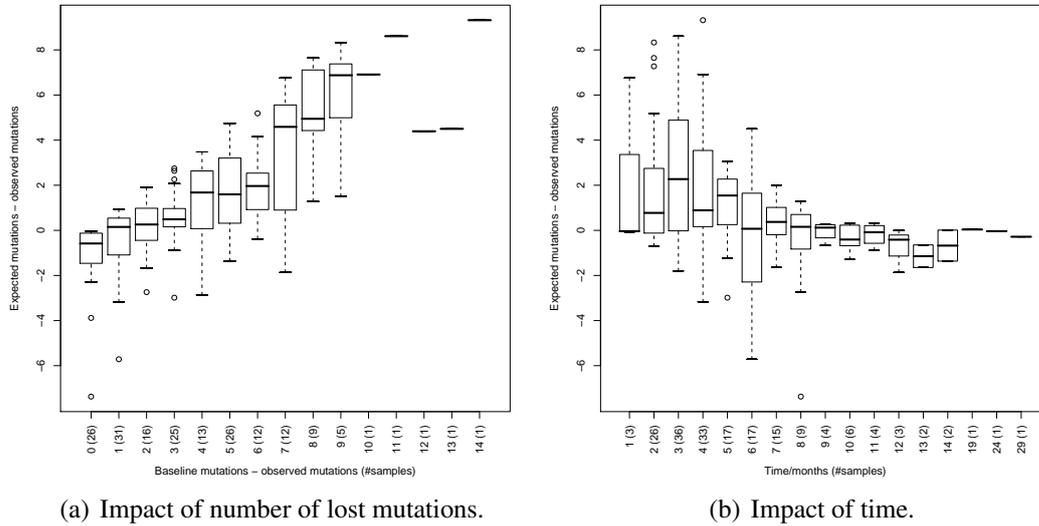


Figure 4.4: Predictors of discordance between the independence model and the observed number of mutations.

tions". In simple regression, both time ( $p = 1.03 \times 10^{-5}$ ) and the actual number of mutations lost ( $p < 2.2 \times 10^{-16}$ ) were highly significantly associated with this difference, whereas baseline viral load ( $p = 0.918$ ) and CD4+ cell count ( $p = 0.8413$ ) were not. A multiple regression model with only “time” and “actual number of mutations lost” as predictors (i.e. no intercept variable) fitted the difference between the expected and observed number of mutations with an  $R^2$  of 0.74, indicating that 74% of the prediction error can be explained only by sampling time and actual number of mutations lost. Both variables were highly significant in the model ( $p < 2 \times 10^{-16}$ ). The coefficients in the model indicate increasing difference with higher number of mutations actually lost, and decreasing difference with increasing time. In other words, the model suggests that the TI episodes that are explained least well by our independence model are those in which a high number of mutations are lost in a very short time. This intuitive characterization can also be seen in [Figure 4.4](#).

### Revertants at RT position 215

The revertants (215CDIVS) provide important indicators for estimating whether the appearance of susceptible variants is more likely due to a mutating resistant population, or due to the outgrowth of an archived wild-type population. A total of 96 out of 144 TIEs started with 215Y (61) or 215F (36) at baseline (1 TIE having a mixture of F and Y at baseline). In 44 of the 61 TIEs with 215Y at baseline (72.1%), Y had disappeared completely at the last follow-up. Likewise, 215F had disappeared completely in 28 of

the 36 TIEs (77.8%). Of the 93 TIEs with either F or Y and no revertants at baseline, revertants are present in the last follow-up in 76 TIEs (81.7%), suggesting that mutation loss in TI takes place primarily (but not exclusively) via mutations in the dominant subpopulation rather than by outgrowth of minority populations with fewer mutations.

#### 4.2.4 Discussion

To our knowledge, this is the first longitudinal study of mutation dynamics based on survival analysis methods, and the largest longitudinal analysis of TI data undertaken to date. The main result of this study is the finding that mutations disappear largely independently from each other in the absence of drug pressure. Despite extensive screening for pair-wise or higher-order interactions among mutations during the disappearance process, no significant dependencies (after correction for multiple testing) were found. This behavior is in stark contrast to the accumulation of resistance mutations in the presence of drug pressure, which is characterized by a clear deviation from independence and the formation of well-pronounced clusters of mutations, as summarized in Section 4.2.1.

We hypothesize the following mechanism for explaining the discordant behavior of mutations in presence versus absence of drug pressure: As summarized in Section 4.2.1, groups of mutations contribute to distinct mechanisms of resistance (improved NRTI recognition, primer unblocking), and it seems natural to expect synergistic or antagonistic behavior of mutations involved in the same or in different mechanisms. In contrast, in the absence of drug, none of the resistance mechanisms fulfills a biological function, and the involved mutations simply become a "nuisance" for the virus. In other words, synergistic or antagonistic interactions are based on a mechanism which is needed for a fitness landscape induced by the presence of drug but not for the natural fitness landscape, which is restored by removing drug pressure. Thus, the mutations disappear independently from each other with little mutual protection. However, we note that our model overestimated the number of remaining mutations in the presence of few baseline mutations, and underestimated it slightly in the presence of many baseline mutations. This apparent deviance from independence could possibly be due to a lack or an abundance of compensatory (fitness-restoring) effects in the former or latter case, respectively. If such a mechanism should exist, it would likely be more pronounced in the case of protease inhibitors, with their substantially larger clusters of secondary mutations.

It is also worth noting that the independence model was least adequate for patients who exhibited a rapid mutation loss within a very short period of time – patients who apparently respond very well to TI. We could not incorporate additional predictors into this model to improve the fit for those patients without compromising the model fit for the majority of patients. Assuming that interactions between viral mutations exhibit consistent effects across subjects, patient-specific genetic factors might be hy-

pothesized as reasons for the difference in mutation dynamics between the majority of patients and the fewer subjects with rapid mutation loss.

We show that in general, therapy interruptions induce a process consisting almost exclusively of resistance mutation loss over time. The appearance during TI of mutations associated with NRTI resistance not present at baseline was a very rare event – in contrast to some previous studies which reported a more frequent appearance (e.g. [Balduin et al., 2005](#)). It is important to consider what “appearance” really means in this context. Given the lack of drug-induced selective pressure during TI, the appearance of a mutation is certainly not to be explained by mutation events within the predominant subpopulation. Rather, it is more likely a population dynamic effect in which a pre-existing but previously undetected viral subpopulation which is also partially resistant to the previous regimen grows over the limit of detection, before wild type variants archived in latent reservoirs take over again in the viral quasispecies. In fact, the rarity of appearance events corresponds well to the high prevalence of revertants at RT position 215, indicating that mutation events are more prominently visible than pure population dynamic effects.

Limitations of this study include the limited number of follow-up samples from each of the patients. Clearly, more densely sampled data (e.g. one sequence per month) on the disappearance of mutations would be desirable. Moreover, strictly speaking, the available sequence data is not simply right-censored, but rather interval-censored, because a mutation can disappear from the main population at any time between two sequencings. However, survival methods for interval-censored data are usually based on rather strong parametric assumptions, in contrast to the non-parametric methods available for right-censored data. The right-censored assumption can also be justified as a conservative scenario of mutation loss, in which mutations disappear just at the last possible time point before the next sampling, and not anywhere in between two samplings. This factor might also contribute to the fact mentioned above that the estimated number of mutations was more often above than below the actually observed number of mutations.

Several other methods for studying viral mutation dynamics, either from cross-sectional or from longitudinal sequence data, have been proposed. In contrast to previously suggested approaches for cross-sectional data, which are based on Bayesian networks ([Deforche et al., 2006, 2007](#)) or mutagenetic trees ([Beerenwinkel et al., 2005a,b](#)), our survival analysis approach makes use of the full information contained in the longitudinal data. Unlike other methods used for longitudinal data (e.g. [Foulkes and DeGruttola, 2003](#); [Beerenwinkel and Drton, 2007](#)), our approach allows for incorporating the effects of baseline covariates (clinical/host parameters, other mutations present at baseline) on mutation dynamics via standard Cox regression. A more sophisticated variation considering the cumulative effect of interactions between mutations during the whole time after onset of TI would require integration of conditional

survival probabilities and would have to allow for transitions between states over time. While such extensions are not necessary to model mutation dynamics during TI (as evidenced by the absence of any sign of significant interaction), it would be necessary to account for the strong interactions among mutations during their accumulation under drug pressure. Another advantage of our approach over previous longitudinal methods is that it provides a natural treatment for the inherently censored nature of the longitudinal sequence data. Finally, it would also be possible to extend the survival approach to allow for both mutation appearance and mutation loss, either by having two different functions for a single mutation, or by using one of the well-developed extensions of classical survival analysis towards recurrent events ([Therneau, 2001](#)).

## 4.3 Improved prediction of drug resistance by evolutionary modeling

Starting with the work of Jaakkola and Haussler, a variety of approaches have been proposed for coupling domain-specific generative models with statistical learning methods. The link is established by a kernel function which provides a similarity measure based inherently on the underlying model. In computational biology, the full promise of this framework has rarely ever been exploited, as most kernels are derived from very generic models, such as sequence profiles or hidden Markov models. In the present section, we introduce the MTreeMix kernel, which is based on a generative model tailored to the underlying biological mechanism. Specifically, the kernel quantifies the similarity of evolutionary escape from antiviral drug pressure between two viral sequence samples. We compare this novel kernel to a standard, evolution-agnostic amino acid encoding in the prediction of HIV drug resistance from genotype, using support vector regression. The results show significant improvements in predictive performance across 17 anti-HIV drugs. Thus, in our study, the generative-discriminative paradigm is key to bridging the gap between population genetic modeling and clinical decision making.

### 4.3.1 Introduction

Kernels provide a general framework of statistical learning that allows for integrating problem-specific background knowledge via the geometry of a feature space. Owing to this unifying characteristic, kernel methods enjoy increasing popularity in many application domains, particularly in computational biology ([Schölkopf et al., 2004](#)). Unfortunately, despite some basic results on the derivation of novel kernels from existing kernels or from more general similarity measures (e.g. via the empirical kernel map ([Schölkopf et al., 2004](#))), the field suffers from a lack of well-characterized design principles. As a consequence, most novel kernels are still developed in an ad hoc

manner.

One of the most promising developments in the recent search for a systematic kernel design methodology is the generative-discriminative paradigm (Jaakkola and Haussler, 1999), also known under the more general term of model-dependent feature extraction (MDFE) (Tsuda et al., 2002a). The central idea of MDFE is to derive kernels from generative probabilistic models of a given process or phenomenon. Starting with Jaakkola and Haussler (Jaakkola and Haussler, 1999) and the seminal work of Amari (Amari and Nagaoka, 2000) on the differential geometric structure of probabilistic models, a number of studies have contributed to an emerging theoretical foundation of MDFE. However, the paradigm is also of immediate intuitive appeal, because mechanistic models of a process that are consistent with observed data and that provide falsifiable predictions often allow for more profound insights than purely discriminative approaches. Moreover, entities that are similar according to a mechanistic model should be expected to exhibit similar behavior in any related properties. From this perspective, MDFE provides a natural bridge between mathematical modeling and statistical learning.

To date, a variety of generic MDFE procedures have been proposed, including the Fisher kernel (Jaakkola and Haussler, 1999) and, more generally, marginalized kernels (Tsuda et al., 2002b), as well as the TOP (Tsuda et al., 2002a), heat (Lafferty and Lebanon, 2003), and probability product kernels (Jebara et al., 2004), along with a number of variations. Surprisingly, however, instantiations of these procedures in bioinformatics have been confined to a very limited number of classical problems, namely protein fold recognition, DNA splice site prediction, exon detection, and phylogenetics. Furthermore, most approaches are based on standard graphical models, such as amino acid sequence profiles or hidden Markov models, that are not adapted in any specific way to the process at hand. For example, a first-order Markov chain along the primary structure of a protein is hardly related to the causal mechanisms underlying polypeptide evolution. Thus, the potential of combining biological modeling with kernelization in the framework of MDFE remains vastly unexplored.

This chapter is motivated by a regression problem from clinical bioinformatics that has recently attracted substantial attention due to its pivotal role in anti-HIV therapy: the prediction of phenotypic drug resistance from viral genotype (reviewed in (Beerewinkel et al., 2005b)). Drug resistant viruses present a major cause of treatment failure and their occurrence renders many of the available drugs ineffective. Therefore, knowing the precise patterns of drug resistance is an important prerequisite for the choice of optimal drug combinations (Clavel and Hance, 2004; Shafer and Schapiro, 2005).

Drug resistance arises as a virus population evolves under partially suppressive antiviral therapy. The extreme evolutionary dynamics of HIV quickly generate viral genetic variants that are selected for their ability to replicate in the presence of the applied drug cocktail. These advantageous mutants eventually outgrow the wild type

population and lead to therapy failure. Thus, the resistance phenotype is determined by the viral genotype. The genotype-phenotype prediction problem is of considerable clinical relevance, because genotyping is much faster and cheaper, while treatment decisions are ultimately based on the viral phenotype (i.e. the level of resistance).

From the perspective of MDFE, the interesting feature of HIV drug resistance lies in the structure of the underlying generative process. The development of resistance involves the stochastic accumulation of mutations in the viral genome along certain mutational pathways. Here, we demonstrate how to exploit this evolutionary structure in genotype-phenotype prediction by deriving a Fisher kernel for mixtures of mutagenetic trees, a family of graphical models designed to represent such genetic accumulation processes. The remainder of this section is organized as follows. In Section 4.3.2, we briefly summarize the mutagenetic trees mixture (MTreeMix) model, originally introduced in (Beerenwinkel et al., 2005b). The Fisher kernel is derived in Section 4.3.3. In Section 4.3.4, the kernel is applied to the genotype-phenotype prediction problem introduced above. We conclude with some of the broader implications of our study, including directions for future work.

### 4.3.2 Background: Mixture models of mutagenetic trees

Consider  $n$  genetic events  $\{1, \dots, n\}$ . With each event  $v$ , we associate the binary random variable  $X_v$ , such that  $\{X_v = 1\}$  indicates the occurrence of  $v$ . In our applications, the set  $\{1, \dots, n\}$  will denote the mutations conferring resistance to a specific anti-HIV drug. Syntactically, a *mutagenetic tree* for  $n$  genetic events is a connected branching  $T = (V, E)$  on the vertices  $V = \{0, 1, \dots, n\}$  and rooted at 0, where  $E \subseteq V \times V$  denotes the edge set of  $T$ . Semantically, the *mutagenetic tree model* induced by  $T$  and the parameter vector  $\theta = (\theta_1, \dots, \theta_n) \in (0, 1)^n$  is the Bayesian network on  $T$  with constrained conditional probability tables of the form

$$\vartheta_v = \begin{matrix} & 0 & 1 \\ \begin{matrix} 0 \\ 1 \end{matrix} & \begin{pmatrix} 1 & 0 \\ 1 - \theta_v & \theta_v \end{pmatrix} \end{matrix}, \quad v = 1, \dots, n.$$

Thus, a mutagenetic tree model is the family of distributions of  $X = (X_1, \dots, X_n)$  that factor as

$$\Pr(X = x \mid \theta) = \prod_{v=1}^n \vartheta_{v, (x_{\text{pa}(v)}, x_v)}.$$

Here,  $x_0 := 1$  (indicating the wild type state without any resistance mutations), and  $\text{pa}(v)$  denotes the parent of vertex  $v$  in  $T$ . Figure 4.5 shows a mutagenetic tree for the development of resistance to the protease inhibitor nelfinavir.

The probability tables impose the constraint that a mutation can only be present if its predecessor in the topology is also present. This restriction sets mutagenetic trees

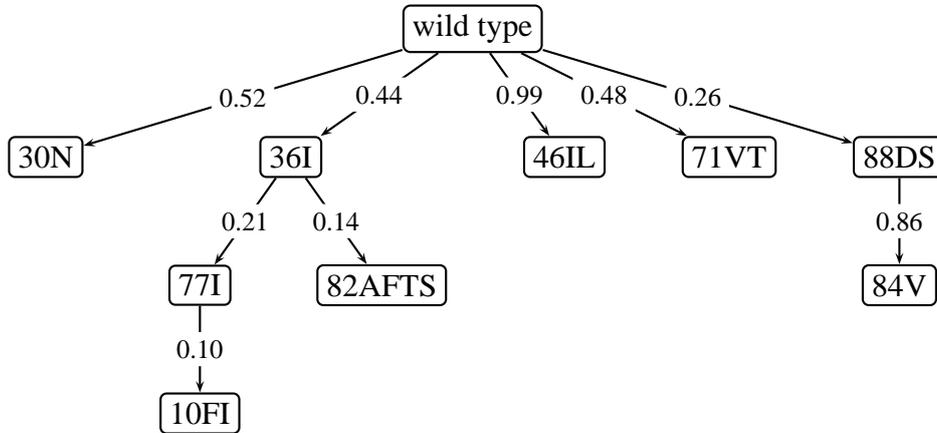


Figure 4.5: Mutagenetic tree for the development of resistance to the HIV protease inhibitor nelfinavir (NFV). Vertices of the tree are labeled with amino acid changes in the protease enzyme. Edges are labeled with conditional probabilities. The tree represents one component of the 6-trees mixture model estimated for this evolutionary process.

apart from standard Bayesian networks in that it allows for an evolutionary interpretation of the tree topology. In particular, the model implies the existence of certain mutational pathways with distinct probabilities. Each pathway is required to respect the order of mutation accumulation that is encoded in the tree. Mutational patterns which do not respect these order constraints have probability zero in the model. We shall exclude these genotypes from the state space of the model. The state space then becomes the following subset of  $\{0, 1\}^n$ ,

$$\mathcal{C} = \{x \in \{0, 1\}^n \mid (x_{\text{pa}(v)}, x_v) \neq (0, 1), \text{ for all } v \in V\},$$

and the factorization of the joint distribution simplifies to

$$\Pr(X = x \mid \theta) = \prod_{\{v \mid x_{\text{pa}(v)}=1\}} \theta_v^{x_v} (1 - \theta_v)^{1-x_v}.$$

The mutational pathway metaphor, originating in the virological literature, is generally considered to be a reasonable approximation to HIV evolution under drug pressure. However, sets of mutational patterns that support different tree topologies are commonly seen in clinical HIV databases. Thus, in order to allow for increased flexibility in modeling evolutionary pathways and to account for noise in the observed data, we consider the larger model class of mixtures of mutagenetic trees. Intuitively, these mixture models correspond to the assumption that a variety of evolutionary forces contribute additively in shaping HIV genetic variability *in vivo*.

Consider  $K$  mutagenetic trees  $T_1, \dots, T_K$  with weights  $\lambda_1, \dots, \lambda_{K-1}$ , and  $\lambda_K = 1 - \sum_{k=1}^{K-1} \lambda_k$ , respectively, such that  $0 \leq \lambda_k \leq 1$  for all  $k = 1, \dots, K$ . Each tree  $T_k$  has parameters  $\theta_k = (\theta_{k,v})_{v=1, \dots, n}$ . The *mutagenetic trees mixture model* is the family of distributions of  $X$  of the form

$$\Pr(X = x \mid \lambda, \theta) = \sum_{k=1}^K \lambda_k \Pr(X = x \mid \theta_k).$$

The state space  $\mathcal{C}$  of this model is the union of the state spaces of the single tree models induced by  $T_1, \dots, T_K$ . In our applications, we will always fix the first tree to be a star, such that  $\mathcal{C} = \{0, 1\}^n$  (i.e., all mutational patterns have non-zero probability). The star accounts for the spontaneous and independent occurrence of genetic events.

### 4.3.3 The MTreeMix Fisher kernel

We now derive a Fisher kernel for the mutagenetic trees mixture models introduced in the previous section. In this paper, our primary motivation is to improve the prediction of drug resistance from viral genotype. However, we defer application-specific details to Section 4, to emphasize the broader applicability of the kernel itself, for example in kernelized principal components analysis.

As Jaakkola and Haussler ([Jaakkola and Haussler, 1999](#)) have suggested, the gradient of the log-likelihood function induced by a generative probabilistic model provides a natural comparison between samples. This is because the partial derivatives in the direction of the model parameters describe how each parameter contributes to the generation of that particular sample. Intuitively, two samples should be considered similar from this perspective, if they influence the likelihood surface in a similar way. The natural inner product for the statistical manifold induced by the log-likelihood gradient is given by the Fisher information matrix ([Amari and Nagaoka, 2000](#)). The computation of this matrix is straightforward, but for practical purposes, the Euclidean dot product  $\langle \cdot, \cdot \rangle$  provides a suitable substitute for the Fisher metric ([Jaakkola and Haussler, 1999](#)).

We first derive the Fisher kernel for the single mutagenetic tree model. The log-likelihood of observing a mutational pattern  $x \in \{0, 1\}^n$  under this model is

$$\ell_x(\theta) = \sum_{\{v \mid x_{\text{pa}(v)}=1\}} x_v \log(\theta_v) + (1 - x_v) \log(1 - \theta_v).$$

Hence, the feature mapping of binary mutational patterns into Euclidean  $n$ -space,

$$\phi : \mathcal{C} \longrightarrow \mathbb{R}^n, \quad x \longmapsto \nabla \ell_x(\theta) = \left( \frac{\partial \ell_x(\theta)}{\partial \theta_1}, \dots, \frac{\partial \ell_x(\theta)}{\partial \theta_n} \right),$$

$x, x'$	$\begin{array}{c} 0 \\ \swarrow \quad \searrow \\ 1 \quad 2 \end{array}$		$\begin{array}{c} 0 \\ \downarrow \\ 1 \\ \downarrow \\ 2 \end{array}$	$\begin{array}{c} 0 \\ \downarrow \\ 2 \\ \downarrow \\ 1 \end{array}$
	00,00	$(\theta_1 - 1)^{-2} + (\theta_2 - 1)^{-2}$	$(\theta_1 - 1)^{-2}$	$(\theta_2 - 1)^{-2}$
00,01	$(\theta_1 - 1)^{-2} + \theta_2^{-1}(\theta_2 - 1)^{-1}$	—	$\theta_2^{-1}(\theta_2 - 1)^{-1}$	
00,10	$\theta_1^{-1}(\theta_1 - 1)^{-1} + (\theta_2 - 1)^{-2}$	$\theta_1^{-1}(\theta_1 - 1)^{-1}$	—	
00,11	$\theta_1^{-1}(\theta_1 - 1)^{-1} + \theta_2(\theta_2 - 1)^{-1}$	$\theta_1^{-1}(\theta_1 - 1)^{-1}$	$\theta_2^{-1}(\theta_2 - 1)^{-1}$	
01,01	$(\theta_1 - 1)^{-2} + \theta_2^{-2}$	—	$(\theta_1 - 1)^{-2} + \theta_2^{-2}$	
01,10	$\theta_1^{-1}(\theta_1 - 1)^{-1} + \theta_2^{-1}(\theta_2 - 1)^{-1}$	—	—	
01,11	$\theta_1^{-1}(\theta_1 - 1)^{-1} + \theta_2^{-2}$	—	$\theta_1^{-1}(\theta_1 - 1)^{-1} + \theta_2^{-2}$	
10,10	$\theta_1^{-2} + (\theta_2 - 1)^{-2}$	$\theta_1^{-2} + (\theta_2 - 1)^{-2}$	—	
10,11	$\theta_1^{-2} + \theta_2^{-1}(\theta_2 - 1)^{-1}$	$\theta_1^{-2} + \theta_2^{-1}(\theta_2 - 1)^{-1}$	—	
11,11	$\theta_1^{-2}\theta_2^{-2}$	$\theta_1^{-2}\theta_2^{-2}$	$\theta_1^{-2}\theta_2^{-2}$	

Table 4.5: Mutagenetic tree Fisher kernels for the three trees on the vertices  $\{0, 1, 2\}$ . The value of the kernel  $K(x, x')$  is displayed for all possible pairs of mutational patterns  $(x, x')$ . Empty cells are indexed with genotypes that are not compatible with the tree.

is given by the Fisher score consisting of the partial derivatives

$$\frac{\partial \ell_x(\theta)}{\partial \theta_w} = \theta_w^{-x_w} (\theta_w - 1)^{x_w - 1} 0^{1 - x_{\text{pa}(w)}} = \begin{cases} \theta_w^{-1}, & \text{if } (x_{\text{pa}(w)}, x_w) = (1, 1) \\ (\theta_w - 1)^{-1}, & \text{if } (x_{\text{pa}(w)}, x_w) = (1, 0) \\ 0, & \text{if } (x_{\text{pa}(w)}, x_w) = (0, 0). \end{cases}$$

Thus, we can define the *mutagenetic tree Fisher kernel* as

$$K(x, x') = \langle \nabla \ell_x(\theta), \nabla \ell_{x'}(\theta) \rangle = \sum_{v=1}^n \theta_v^{-(x_v + x'_v)} (\theta_v - 1)^{(x_v + x'_v) - 2} 0^{2 - (x_{\text{pa}(v)} + x'_{\text{pa}(v)})}.$$

For example, the Fisher kernels for the three mutagenetic trees on  $n = 2$  genetic events are displayed in [Table 4.5](#).

To better understand the operation of the novel kernel, we rewrite the kernel function  $K$  as follows:

$$K(x, x') = \sum_{v=1}^n \kappa(\theta_v)_{(x_{\text{pa}(v)}, x_v), (x'_{\text{pa}(v)}, x'_v)},$$

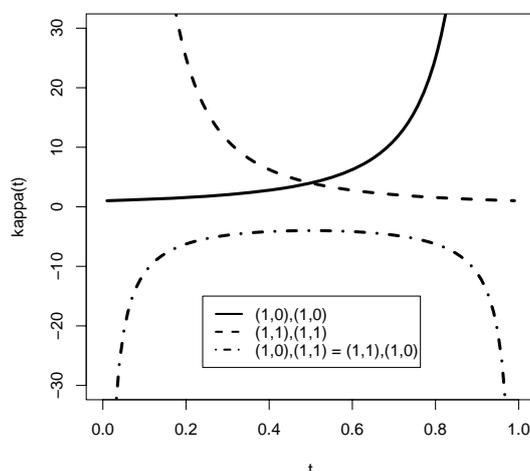


Figure 4.6: Non-zero entries of the matrix  $\kappa(t)$  that defines the mutagenetic tree Fisher kernel. The three graphs are indexed in the same way as the matrix, namely by pairs  $((x_{\text{pa}(v)}, x_v), (x'_{\text{pa}(v)}, x'_v))$  denoting the value of two genotypes  $x$  and  $x'$  at an edge  $(\text{pa}(v), v)$  of the mutagenetic tree. The graphs illustrate that the largest contributions stem from shared, unlikely mutations (positive effect, solid and dashed line) and from differing, likely or unlikely mutations (negative effect, dash-dot line).

with  $\kappa$  defined as

$$\kappa(t) = \begin{matrix} & (0, 0) & (1, 0) & (1, 1) \\ \begin{matrix} (0, 0) \\ (1, 0) \\ (1, 1) \end{matrix} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & (t-1)^{-2} & t^{-1}(t-1)^{-1} \\ 0 & t^{-1}(t-1)^{-1} & t^{-2} \end{pmatrix} \end{matrix}$$

The matrix  $\kappa(t)$  is indexed by pairs of pairs  $((x_{\text{pa}(v)}, x_v), (x'_{\text{pa}(v)}, x'_v))$ . The non-zero entries of  $\kappa$  are displayed in Figure 4.6 as functions of the parameter  $t$ . An edge contributes strongly to the kernel value, if the two genotypes agree on it, but the common event (occurrence or non-occurrence of the mutation) was unlikely (Figure 4.6, solid and dashed line). If the two genotypes disagree, the edge contributes negatively, especially for extreme parameters  $\theta_v$  close to zero or one (Figure 4.6, dash-dot line), which make one of the events very likely and the other very unlikely. Thus, the application of the Fisher kernel idea to mutagenetic trees leads to a kernel that measures similarity of evolutionary escape in a way that corresponds well to virological intuition.

Due to the linear mixing process, extending the Fisher kernel from a single mutagenetic tree to a mixture model is straightforward. Let  $\ell_x(\lambda, \theta) = \log \Pr(x \mid \lambda, \theta)$  be

the log-likelihood function, and denote by

$$\gamma_l(x | \lambda, \theta) = \frac{\lambda_l \Pr(x | \theta_l)}{\Pr(x | \lambda, \theta)}$$

the responsibility of tree component  $T_l$  for the observation  $x$ . Then the partial derivatives with respect to  $\theta$  can be expressed in terms of the partials obtained for the single tree models, weighted by the responsibilities of the trees,

$$\frac{\partial \ell_x(\lambda, \theta)}{\partial \theta_{l,w}} = \gamma_l(x | \lambda, \theta) \frac{\partial \ell_x(\theta_l)}{\partial \theta_{l,w}}.$$

Differentiation with respect to  $\lambda$  yields

$$\frac{\partial \ell_x(\lambda, \theta)}{\partial \lambda_l} = \frac{\Pr(x | \theta_l) - \Pr(x | \theta_K)}{\Pr(x | \lambda, \theta)}.$$

We obtain the *mutagenetic trees mixture (MTreeMix) Fisher kernel*

$$\begin{aligned} K(x, x') &= \langle \nabla \ell_x(\lambda, \theta), \nabla \ell_{x'}(\lambda, \theta) \rangle \\ &= \sum_{l=1}^{K-1} \frac{[\Pr(x | \theta_l) - \Pr(x | \theta_K)] [\Pr(x' | \theta_l) - \Pr(x' | \theta_K)]}{\Pr(x | \lambda, \theta) \Pr(x' | \lambda, \theta)} \\ &\quad + \sum_{l=1}^K \sum_{w=1}^n \gamma_l(x | \lambda, \theta) \gamma_l(x' | \lambda, \theta) \kappa(\theta_{l,w})_{(x_{\text{pa}(w)}, x_w), (x'_{\text{pa}(w)}, x'_w)}. \end{aligned}$$

### 4.3.4 Experimental results

In this section, we use the Fisher kernel derived from mutagenetic tree mixtures for predicting HIV drug resistance from viral genotype. Briefly, resistance is the ability of a virus to replicate in the presence of drug. The degree of resistance is usually communicated as a non-negative number. This number indicates the fold-change increase in drug concentration that is necessary to inhibit viral replication by 50%, as compared to a fully susceptible reference virus. Thus, higher fold-changes correspond to increasing levels of resistance. We consider all fold-change values on a  $\log_{10}$  scale.

Information on phenotypic resistance strongly affects treatment decisions, but the experimental procedures are too expensive and time-consuming for routine clinical diagnostics. Instead, at the time of therapy failure, the genotypic makeup of the viral population is determined using standard sequencing methods, leaving the challenge of inferring the phenotypic implications from the observed genotypic alterations. It is also desirable to minimize the number of sequence positions required for reliable determination of drug resistance. With a small number of positions, sequencing could be replaced by the much cheaper line-probe assay (LiPA) technology (Schmit et al.,

1998), which focuses on the determination of mutations at a limited number of pre-selected sites. This method could bring resistance testing to resource-poor settings in which DNA sequencing is not affordable.

All approaches to this problem described to date are based on a direct correlation between genotype and phenotype, without any further modelling involved. Application of the Fisher kernel to this task is motivated by the hypothesis that the traces of evolution present in the data and modelled by mutagenetic trees mixture models can provide additional information, leading to improved predictive performance. In a recent comparison of several statistical learning methods, support vector regression attained the highest average predictive performance across all drugs (Rabinowitz et al., 2006). Accordingly, we have chosen this best-performing method to compare to the novel kernel.

Specifically, our experimental setup is as follows. For each drug, we start with a genotype-phenotype data set (Walter et al., 1999) of size 305 to 858 (Table 4.6, column 3). Based on a list of resistance mutations maintained by the International AIDS Society (Johnson et al., 2005), we extract the residues listed in column 2. The number indicates the position in the viral enzyme (reverse transcriptase for the first two groups of drugs, and protease for the third group), and the amino acids following the number denote the mutations at the respective site that are considered resistance-associated. For example, the feature vector for the drug zidovudine (ZDV) consists of six variables representing the reverse transcriptase mutations 41L, 67N, 70R, 210W, 215F or Y, and 219E or Q. In the naive indicator representation, a mutational pattern within these six mutations is transformed to a binary vector of length six, each entry encoding the presence or absence of the respective mutation.

The Fisher kernel requires a mutagenetic trees mixture model for each of the evaluated drugs. Using the MTreeMix software package<sup>6</sup>, these models were estimated from an independent set of sequences derived from patients failing a therapy that contained the specific drug of interest. In 100 replicates of ten-fold cross-validation for each drug model, we then recorded the squared correlation coefficient ( $r^2$ ) of indicator variable-based versus Fisher kernel-based support vector regression. Avoiding both costly double cross-validation with the limited amount of data and overfitting with single cross-validation, we fixed standard parameters for both SVMs. As suggested by Jaakkola and Haussler (Jaakkola and Haussler, 1999), the Fisher kernel may be combined with additional transformations. Thus, we evaluated the standard kernels for both setups. For the indicator representation, the linear kernel performed best, whereas the Fisher scores performed best when combined with a Gaussian RBF kernel. We used these two kernels in the final comparison reported in Table 4.6.

The results displayed in columns 5 and 6 of Table 4.6 show the improvements attained via the Fisher kernel method as estimated by the squared correlation coefficient,

---

<sup>6</sup><http://mtreemix.bioinf.mpi-sb.mpg.de>

$r^2$ . After correction for multiple comparisons, the null hypothesis of equal mean was rejected ( $P < 0.01$ , Wilcoxon test) in 15 out of 17 cases, a ratio that is highly unlikely to occur by chance ( $P < 0.0025$ , binomial test). The most drastic improvements were obtained for the drugs 3TC, NVP and NFV. Slight decreases were observed for ddC and APV. Interestingly, when we combined both feature vectors, the cross-validated performance of the combined predictor was consistently at least as good as the best individual predictor (data not shown). We obtained similar results when evaluating performance by the mean squared error instead of the correlation coefficient (data not shown).

### 4.3.5 Conclusions

The Fisher kernel derived in this paper allows for leveraging stochastic models of HIV evolution in many kernel-based scenarios. To our knowledge, this is the first study in which a probabilistic model tailored to a specific biological mechanism (namely, the evolution of drug resistance) is exploited in a discriminative context. Using the example of inferring drug resistance from viral genotype, we showed that significant improvements in predictive performance can be obtained for almost all currently available antiretroviral drugs. These results provide strong incentive for further exploitation of evolutionary models in clinical decision making. Moreover, they also underline the potential benefits from integrating several sources of data (genotype-phenotype, evolutionary). The high correlation that can be observed with a relatively small number of mutations was unexpected and suggests that reliable resistance predictions can also be obtained on the basis of LiPA assays which are much cheaper than standard sequencing technologies. While our choice of mutations was based on a selection from the literature, an interesting problem would be to design dedicated LiPA assays containing a set of mutations that allow for optimal prediction performance in this generative-discriminative setting. Finally, mixtures of mutagenetic trees have already been applied in other contexts, for example to model progressive chromosomal alterations in cancer ([Rahnenfuhrer et al., 2005](#)), and we expect kernel methods to play an important role in this context, too.

### 4.3.6 Acknowledgment

The work reported in this section was performed in collaboration with Niko Beerenwinkel (Berkeley).

Table 4.6: Comparison of support vector regression performance for the MTreeMix Fisher kernel ( $F$ ) versus a naive amino acid indicator ( $I$ ) representation. The drugs (first column) are grouped into the three classes of nucleoside/nucleotide reverse transcriptase inhibitors (rows 1–7), nonnucleoside reverse transcriptase inhibitors (rows 8–10), and protease inhibitors (rows 11–17). MTreeMix models were estimated based on the mutations listed in the second column. The third column indicates the number  $N$  of available genotype-phenotype pairs, and the number  $K$  of trees in the mixture model is shown in column 4. Columns 5 and 6 indicate the squared correlation coefficients, averaged across 100 replicates of 10-fold cross-validation. P-values (last column) are obtained from Wilcoxon rank sum tests (for the null hypothesis that the location of the distributions of  $r_F^2$  and  $r_I^2$  are identical), correcting for multiple testing using the Benjamini-Hochberg method.

DRUG	MUTATIONS	N	K	$r_F^2$	$r_I^2$	$\log_{10} P$
ZDV	41L, 67N, 70R, 210W, 215FY, 219EQ	856	5	0.61	0.57	$< -15.0$
3TC	44D, 118I, 184IV	817	5	0.71	0.64	$< -15.0$
ddI	65R, 67N, 70R, 74V, 184V, 210W, 215FY, 219EQ	858	4	0.28	0.24	$< -15.0$
ddC	41L, 65R, 67N, 70R, 74V, 184V	536	2	0.25	0.26	$-0.3$
d4T	41L, 67N, 70R, 75TMSA, 210W, 215YF, 219QE	857	4	0.22	0.21	$-2.7$
ABC	41L, 65R, 67N, 70R, 74V, 115F, 184V, 210W, 215YF	846	7	0.57	0.55	$-9.0$
TDF	41L, 65R, 67N, 70R, 210W, 215YF, 219QE	527	3	0.45	0.43	$-7.0$
NVP	100I, 103N, 106A, 108I, 181CI, 188CLH, 190A	857	5	0.58	0.49	$< -15.0$
EFV	100I, 103N, 108I, 181CI, 188L, 190SA	843	4	0.60	0.56	$< -15.0$
DLV	103N, 181C	856	2	0.49	0.48	$-1.7$
IDV	10IRV, 20MR, 24I, 32I, 36I, 46IL, 54V, 71VT, 73SA, 77I, 82AFT, 84V, 90M	851	4	0.65	0.63	$-14.3$
SQV	10IRV, 48V, 54VL, 71VT, 73S, 77I, 82A, 84V, 90M	854	4	0.68	0.66	$-8.6$
RTV	10FIRV, 20MR, 24I, 32I, 33F, 36I, 46IL, 54VL, 71VT, 77I, 82AFTS, 84V, 90M	855	4	0.77	0.75	$-12.0$
NFV	10FI, 30N, 36I, 46IL, 71VT, 77I, 82AFTS, 84V, 88DS	853	6	0.62	0.55	$< -15.0$
APV	10FIRV, 32I, 46IL, 47V, 50V, 54LVM, 73S, 84V, 90M	665	3	0.58	0.59	$-2.0$
LPV	10FIRV, 20MR, 24I, 32I, 33F, 46IL, 47V, 50V, 53L, 54LV, 63P, 71VT, 73S, 82AFTS, 84V, 90M	507	5	0.73	0.69	$< -15.0$
ATV	32I, 46I, 50L, 54L, 71V, 73S, 82A, 84V, 88S, 90M	305	2	0.54	0.52	$-2.4$



# Drug Resistance: Genotype, Phenotype, and Virological Response

By using evolutionary modeling for predicting drug resistance, Section 4.3 has provided a bridge from viral evolution to drug resistance, whose various manifestations – genotypic, phenotypic, clinical – will be the subject of the present chapter. We start (Section 5.1) with a comprehensive review of the algorithms that have been proposed to date for relating genotypic alterations to changes in the viral phenotype or in the virological response to therapy. We then describe novel approaches for identifying resistance mutations and for characterizing their co-occurrence behavior, followed by an application to nucleoside and non-nucleoside reverse transcriptase inhibitors (Section 5.2). The question of whether clinical resistance should be predicted by direct correlation of genotypes with virological response, or by using predicted phenotypes as an “intermediate” step has been subject to some controversy. We study this question in a large-scale evaluation in Section 5.3. Concluding the chapter, we investigate potential benefits of incorporating an additional phenotypic property of the virus – “fitness”, or, more accurately, replication capacity – into models for virological response prediction (Section 5.4).

## 5.1 Algorithms for the interpretation of drug resistance: a review

### 5.1.1 Introduction: selection of antiretroviral therapy

Physicians treating HIV-1-infected patients are faced with selecting an optimal new regimen upon therapy failure. This task is highly complex because of the increasing number of available antiretroviral drugs, significant cross-resistance and the likely

presence of archived drug-resistant viral variants selected by previous regimens. Parameters with potential impact on treatment decisions include the plasma viral load, CD4<sup>+</sup> cell count, viral genotype, phenotype and replication capacity, and pharmacological data. Other factors to consider include tolerability, toxicity and the ability to preserve future treatment options.

Interpretation algorithms are designed to assist the treating physician in choosing an optimal drug combination using information from drug-resistance testing. In this context, ‘interpretation’ refers to the task of predicting a specific factor (i.e. drug activity or virological response) from one or more other factors. The word ‘algorithm’ originates from computer science and denotes a set of well-defined instructions for accomplishing a given task. In clinical practice, the most popular interpretation algorithms are rule-based approaches for predicting drug activity from the viral genotype. Published reviews provide in-depth coverage of rule-based approaches ([Schmidt et al., 2002](#); [Zolopa et al., 2004](#)). However, the field has broadened considerably in scope and methodology, and new tools are currently being developed.

### 5.1.2 The factors of interest: drug activity, viral phenotype and virological response

Interpretation algorithms vary considerably in goals and methodology. However, due to the lack of a commonly accepted terminology, they are often collectively referred to as ‘resistance algorithms’ or ‘resistance scores’. To avoid confusion, this chapter distinguishes between algorithms for predicting *in vivo* antiviral activity of a single drug, *in vitro* viral phenotypic resistance or replication capacity, and *in vivo* virological response to a combination of drugs. From the clinical point of view, we are ultimately interested in the latter (measured as the decline in plasma viral load). However, drug activity and viral phenotype can be valuable intermediates in predicting responses to combination therapy (cf. also [Section 5.3](#)). For an appropriate use of interpretation algorithms, it is crucial to understand the differences between these three key terms.

Phenotypic drug resistance and replication capacity refer to the ability of a virus to replicate in the presence and absence of drug, respectively. These parameters are measured using controlled and reproducible experimental assays. Results of phenotypic assays are typically reported as the fold change (FC) in the drug concentration that inhibits viral growth by 50% ( $IC_{50}$ ) relative to a control wild-type virus for each drug. The replication capacity is reported as a percentage relative to the reference strain. It should be kept in mind that the ability of a virus to replicate decreases continually with increasing drug concentration, and that measurements of FC and replication capacity can be used to summarize this dose-response curve.

Replication capacity assays have only recently been introduced and are yet to gain a firm place in clinical practice; however, assays for measuring phenotypic resistance

have been in use for about 10 years and are able to predict virological response (DeGruttola et al., 2000). Nevertheless, this association is not perfect, and several cases of discordance between FC values and virological response *in vivo* have been reported. For example, the reverse transcriptase (RT) mutation M184V is associated with resistance to didanosine (ddI) and abacavir (ABC), but in isolation has little or no discernible impact on virological response to these drugs (Brun-Vézinet et al., 2004). Similarly, patients who develop M184V during lamivudine (3TC) monotherapy maintain a plasma viral load of approximately  $0.5 \log_{10}$  copies/ml lower than at baseline, despite maximal levels of phenotypic resistance to the drug (Eron et al., 1995). Based on these observations, the clinical use of phenotypic resistance data has been said to suffer from ‘serious limitations as the association between *in vitro* resistance levels and virological response is often not well characterized’ (Brun-Vézinet et al., 2004). However, it is conceivable that these apparent shortcomings can be alleviated by taking into account additional features of the dose-response curve. For example, replication capacity assays show a marked replicative defect associated with the M184V mutation (Devereux et al., 2001), thus providing a possible explanation for the observed virological suppression despite phenotypic resistance. In summary, phenotypic drug-resistance data alone cannot explain the full range of *in vivo* virological response effects.

The concept of *in vivo* antiviral activity of a single drug is slightly more abstract than *in vitro* phenotype and *in vivo* virological response. The objective of this measurement is to capture information about the antiviral effect of a single drug. However, anti-HIV drugs are generally given in combination, and synergistic and antagonistic effects between the drugs in a regimen play a significant role in determining virological response; an up-to-date review can be found in (Boffito et al., 2005), and the University of Liverpool provides an informative on-line resource<sup>1</sup>. Thus, it is conceptually problematic to evaluate in isolation a model of single-drug activity because the effects induced by the drug are not independent from the effects of co-administered drugs. Consequently, models of drug activity are often evaluated in the context of combination therapy, where an additional algorithm is used to combine the activity of individual drugs into a score for the drug combination. While this suggests a straightforward evaluation procedure, it is important to emphasize that it is the integration of a drug activity model and an activity combination step that is evaluated here, and not an activity model in isolation. Nevertheless, different activity models can be compared by keeping the combination algorithm fixed. To summarize, the concept of *in vivo* drug activity has proven to be a useful intermediate between *in vitro* measurements and the virological response to a combined regimen.

---

<sup>1</sup>[www.hiv-druginteractions.org](http://www.hiv-druginteractions.org)

### 5.1.3 Predicting *in vivo* antiviral activity of individual drugs

Algorithms for predicting *in vivo* drug activity can be either derived from automatic, data-driven statistical procedures or developed by panels of experts. They can be formulated as simple tables or as complex computational procedures. Finally, they can be based on genotype, phenotype, or genotype with predicted phenotype as an intermediate step, and can also take into account additional relevant factors, such as pharmacokinetic data.

#### Predicting *in vivo* drug activity from the viral genotype

As genotyping is used widely, algorithms for predicting *in vivo* activity from the genotype provide valuable tools to support decision-making in clinical practice. The main limitations in relating the viral genotype to drug activity include paucity of monotherapy data and the diversity and complexity of mutational patterns. Currently, more than 80 mutations have been implicated in drug resistance *in vitro* or reduced drug activity *in vivo* (Johnson et al., 2005), and the number is likely to grow as more patients are exposed to existing agents in varying combinations (Gonzales et al., 2003; Wu et al., 2003; Svicher et al., 2006), increasing attention is devoted to non-B HIV-1 subtypes (Camacho, 2006), new compounds are introduced and new resistance mutations are identified from large databases.

The most popular genotype interpretation algorithms include:

- **ANRS AC 11**, Version 09/2005 (Brun-Vézinet et al., 2003). The algorithm specification is available at [www.hivfrenchresistance.org](http://www.hivfrenchresistance.org). Free web-based services based on this algorithm are available at [pugliese.club.fr/resistance.html](http://pugliese.club.fr/resistance.html) (in French) and [hivdb.stanford.edu/pages/algs/HIValg.html](http://hivdb.stanford.edu/pages/algs/HIValg.html). There are two output categories for NRTIs and PIs, and three output categories for NNRTIs and FIs.
- **GuideLines**, Version 9.0 (Bayer Diagnostics). This proprietary algorithm accompanies the TruGene Genotyping System. Information is available at [www.labnews.de](http://www.labnews.de).
- **HIVdb**, Version 4.1.2 (05/2005). The algorithm is available at [hivdb6.stanford.edu/asi/deployed/HIVdb.html](http://hivdb6.stanford.edu/asi/deployed/HIVdb.html). It has five output categories, derived from a more fine-grained scoring system.
- **HIVgrade**, Version 1.0. A free web-based service is available at [www.hiv-grade.de](http://www.hiv-grade.de). The algorithm has four output categories.
- **Rega**, Version 6.4 (Van Laethem et al., 2002). The algorithm specification is available at [www.kuleuven.be/rega/cev/pdf/](http://www.kuleuven.be/rega/cev/pdf/)

RegaResistanceAlgorithmv6.4.0\_1Dec2004.pdf. A free web-based service is available at [hivdb6.stanford.edu/asi/deployed/HIValg.html](http://hivdb6.stanford.edu/asi/deployed/HIValg.html).

- **Retrogram**, Version 1.6i (Virology Networks). The algorithm is available at [www.retrogram.com](http://www.retrogram.com). It has four output categories, plus one category “no prediction” (due to insufficient data).
- **ViroSeq**, Version 2.6 (Celera Diagnostics). This proprietary algorithm accompanies the ViroSeq HIV-1 Genotyping System (Abbott). Information is available at [www.celeradiagnostics.com/cdx/ViroSeq](http://www.celeradiagnostics.com/cdx/ViroSeq). There are three output categories.
- **geno2pheno**, Version 3.0 (05/2005), ([Beerenwinkel et al., 2002, 2003a](#)). A free web-based service for predicting phenotypic resistance is available at [www.geno2pheno.org](http://www.geno2pheno.org). Predictions offered include a regression of the fold-change in  $IC_{50}$  and a classification with two output categories.
- **VircoType** (Virco BVBA). This proprietary algorithm predicts fold-change in  $IC_{50}$ , and a three-category output based on clinical cutoffs.
- **THEO**, Version 1.0 (04/2005), ([Altmann et al., 2007](#)). A free web-based service for ranking combination therapies is available at [www.geno2pheno.org](http://www.geno2pheno.org). The ranking criterion is the probability that a given patient will respond to a given drug combination.

The algorithms ANRS, GuideLines, HIVdb, HIVgrade, Rega, Retrogram, and ViroSeq aim at predicting *in vivo* activity from genotype, while geno2pheno and VircoType predict *in vitro* (phenotypic) resistance, although the VircoType also includes a prediction of *in vivo* activity based on the predicted phenotypic resistance and clinical cutoffs. Finally, THEO aims at predicting virological response to combination therapy. The remainder of this section will briefly compare the scoring schemes, reported levels of drug activity, and the design and evaluation of these different approaches.

**Scoring schemes.** At the heart of most genotype interpretation algorithms are simple scoring schemes, based on the presence or absence of specific mutations. For example, one of the rules for predicting resistance to ddI in the current version of the Rega algorithm is based on the scheme [K65R+T69G/N+”T69 insertion”+L74V+Q151M]. Thus, a genotype with mutation K65R and L74V would be assigned a score of 2. Scores are often turned into a fixed set of activity levels (such as active versus non-active) via the choice of a cut-off (see below). Our example would be classified as ‘ddI inactive’ by the Rega algorithm because the scheme comes with a cut-off of 1.

Scores may also take into account resensitization effects induced by specific mutations. For example, one of the ANRS rules for predicting ddI resistance is based on the scheme [M41L+T69D+L74V+T215Y/F+K219E/QK70RM184V/I], with the resensitization condition that the presence of K70R or M184V/I leads to a subtraction from the score.

In reality, not all mutations contribute equally to reduced activity of a drug. The HIVdb algorithm accommodates this observation by assigning a weight (expert-derived) to each mutation. Again, negative weights indicate resensitization effects of varying strength. For example, the current HIVdb ddI scoring scheme is based on 51 mutations with weights ranging from +2 to +50. When limited to the mutations given in the REGA score mentioned above, the HIVdb score reads as follows: [30×K65R+20×T69G+40×T69 insertion+50×L74V+50×Q151M]. Non-linear interactions between specific mutations are usually ignored in these models.

**Levels of drug activity.** Although drug activity is a continuum, for ease of usage, scores are typically split into a small number of drug activity levels. The most popular approaches employ three classes of drug activity: active, inactive and intermediately active. Unfortunately, most activity algorithms use the terms susceptible and resistant which should be restricted to approaches for predicting *in vitro* drug resistance. For example, the ANRS score for tenofovir [M41L+E44D+D67N+T69D/N/S+L74V+L210W+T215F/Y] is categorized using the genotypic cut-offs ' $\leq 3$  mutations' (active), ' $3-5$  mutations' (partially active) and ' $\geq 6$  mutations' (inactive). The ANRS algorithm assigns only two levels of activity to non-nucleoside reverse transcriptase inhibitors (NNRTIs) and the fusion inhibitor enfuvirtide, in contrast to the situation with nucleoside reverse transcriptase inhibitors (NRTIs) and protease inhibitors (PIs). Other approaches that use three activity levels include GuideLines, Rega and ViroSeq. Four levels are used by HIV-GRADe and RetroGram (which have an additional special class 'insufficient data' for refraining from a prediction), whereas HIVdb has five distinct activity levels. Although these categories of activity may facilitate interpretation, the use of raw scores may be useful when predicting response to drug combinations (Swanstrom et al., 2004).

**Design and evaluation.** Statistical approaches for deriving drug activity from genotype and treatment response data are summarised in (Brun-Vézinet et al., 2004). Crucial decisions include the way the response is measured and the time of analysis. Typically, relevant mutations are first identified by using a feature-selection method, such as simple univariate statistical tests, followed by a multivariate model-building procedure. This approach has been used for example in the ANRS Narval study (Brun-Vézinet et al., 2003). However, most methods are not based on statistical analysis, but rather on consensus from expert panels based on current literature. Several compar-

ative evaluations of these models have been performed (De Luca et al., 2003, 2004), and further initiatives are on the way (Costagliola et al., 2005).

#### 5.1.4 Predicting *in vivo* drug activity from the viral phenotype

At first sight, predicting drug activity from FC values obtained in phenotypic assays may seem easier than interpreting genotypic results. However, despite this apparent simplicity, there is no gold standard for interpreting phenotypic results. The bulk of work on this topic has focused on the choice of the appropriate FC cut-offs to categorize drugs into ‘active’ or ‘non-active’. The use of cut-offs is based on the assumption that drugs have a binary activity profile, being completely active below, and completely inactive above, a given level of phenotypic resistance. However, for most antiretroviral drugs, the relation between phenotypic resistance and *in vivo* activity is not an all-or-nothing phenomenon. Distinct levels of activity can be derived from FC data by setting two cut-offs, one for diminished activity (the lower cut-off) and one for abolished activity (the higher cut-off). For example, the VircoTYPE system (Virco BVBA) has recently adopted two cut-offs for each drug, categorizing the continuous drug activity as full, reduced by 20% or reduced by 80% (Bachelier et al., 2004).

Remarkably, recent studies indicate that the use of models based on cut-offs does not necessarily improve prediction of virological response. Swanstrom et al. (2004) compared activity derived from a single cut-off with a continuous activity score in which FC values below 2.5 and above 10 were mapped to 0 (completely active) and 1 (completely inactive), respectively, with linear interpolation in between. The model using continuous activity score, but not that with cut-offs, was significantly associated with virological response.

An interesting combination of a dichotomous model with a continuous score is described by Beerenwinkel et al. (2003b). Based on the observation that the distribution of predicted resistance factors closely follows a mixture of two Gaussians (the two components accounting for susceptible and resistant samples), activity of a drug is defined as the probability that the corresponding FC value belongs to the susceptible population.

#### 5.1.5 Predicting *in vitro* resistance and replication capacity from the viral genotype

Phenotypic assays are more expensive and laborious than genotypic assays, and thus less suitable for routine clinical use. There is also evidence that phenotype predictions derived from genotypic data might be more accurate than a single phenotypic measurement as a result of the inherent variability of phenotypic assays (Van Houtte et al., 2004). Moreover, genotypic testing can detect the presence of resistance mutations

that signal emerging resistance before a significant effect on phenotype is measurable *in vitro*. For these reasons, a considerable amount of work has been devoted to developing systems for accurately predicting the viral phenotype from genotypic data.

To date, only two comprehensive systems are available, the VircoType provided by Virco BVBA and geno2pheno (Beerenwinkel et al., 2002, 2003a), which is freely available. Both systems are based on statistical learning methodology (Hastie et al., 2001), which constructs predictive models from 'training' data in the form of matched genotype-phenotype pairs. Despite this common background, the models are very different in nature: the VircoType is derived from an instance-based method, also known as nearest-neighbour learning. From a query sequence, a drug-specific mutational profile is derived, based on predefined sequence positions. All samples with this profile are then retrieved from a large database of experimentally determined genotype-phenotype pairs. The predicted phenotype is the mean FC of these matched samples. geno2pheno is based on a much smaller database ( $\approx 900$  samples per drug) than the VircoType ( $\approx 25,000$  samples per drug). Whereas the latter is essentially model-free, relying only on averaging experimentally determined phenotypes, geno2pheno is based on a mathematical model known as the support vector machine (SVM). In the training step, sequences with known phenotype are mapped into a high-dimensional vector space. In this space, a hyperplane is computed which optimally approximates the genotype-phenotype relation. The final SVM models of geno2pheno are high-dimensional linear models that assign weights to individual mutations, which are then added to yield the predicted FC, similar to in the HIVdb model. The VircoType has been shown to be superior to standard phenotyping (when using a single measurement), whereas the prediction for geno2pheno is highly correlated with phenotypic resistance for most drugs.

Other approaches have been considered based on a variety of statistical learning approaches, including linear discriminant analysis (Sevin et al., 2000), linear regression with (Vermeiren et al., 2004) or without interaction terms (Wang et al., 2004a), a novel non-parametric statistical method (DiRienzo et al., 2003), decision trees (Beerenwinkel et al., 2002) and artificial neural networks (Wang and Larder, 2003), in addition to a semi-supervised approach based on self-organising maps using predicted structural features of the enzyme-ligand complex (Draghici and Potter, 2003).

All approaches based on statistics or statistical learning crucially rely on 'training' data in the form of matched genotype-phenotype pairs. However, such data are hard to obtain for novel drugs, thus delaying the development of genotype-phenotype models. A possible approach to deriving genotype-phenotype models without training data is structure-based phenotyping. This method is based on the assumption that resistance can be predicted from computing changes in binding energy using molecular modelling. The only prerequisite is the availability of at least one crystal structure of the target molecule, which is then adapted computationally to specific mutant genotypes.

For example, Cengent Therapeutics and Quest Diagnostics have developed structure-based phenotyping models for six PIs and, in comparing the models to the PhenoSense or Antivirogram assays, report squared correlation coefficients between 0.34 and 0.83, depending on drug and data source. Whereas the correlation is lower than that of the data-driven approaches, structure-based phenotyping may be useful at stages when genotype-phenotype data are scarce. Intrinsic drawbacks include the dependence on an adequate crystal structure (which can be problematic in the case of the envelope glycoprotein gp120 or the transmembrane glycoprotein gp41) and the inability to capture effects that take place across large distances in the enzyme or that happen far from the active site.

As mentioned above, drug activity is a function of multiple factors, including resistance. One of these factors, which can also be measured *in vitro*, is the viral replication capacity. Models for predicting activity from FC and replication capacity might be superior to approaches based on FC alone (but cf. Section 5.4). Thus, as in the case of phenotypic resistance, approaches for predicting replication capacity from genotype may prove useful in routine clinical practice. To date, two approaches have been suggested. Segal et al. (2004) have used decision trees and random forests, and a custom ‘deletion/substitution/addition’ regression algorithm has been developed by Birkner et al. (2005). In both models, the M184V mutation was predictably found to be associated with reduced replication capacity, along with a number of other mutations. A comprehensive screening for protease and RT positions associated with changes in replication capacity has been performed (Bonhoeffer et al., 2004), based on almost 10,000 matched genotype-replication capacity pairs.

### 5.1.6 Predicting *in vivo* virological response to combination therapy

The most common approach to predicting virological response to a combination of drugs proceeds in two steps. The activities of the individual drugs in the regimen are predicted using one of the models described above, and combined into an overall response score for the whole regimen. One method simply sums up all predicted activities for the drugs in the regimen. Depending on the data source, this approach leads to the calculation of a genotypic (GSS) or phenotypic (PSS) *susceptibility score* (DeGruttola et al., 2000). For example, given a viral population with the RT mutation K65R and the protease mutations V32I and I47A, when using the ANRS algorithm for predicting activity, a regimen comprising ABC, 3TC and ritonavir-boosted lopinavir would be assigned a GSS of  $0.5+0.5+0=1$ , whereas the combination of zidovudine (ZDV), ddI and efavirenz (EFV) would attain a score of  $1+0.5+1=2.5$ . Likewise, given phenotypic measurements, scores of 0 and 1 (or intermediate scores) are assigned based on FC cut-offs. Susceptibility scores can also be weighted, for example with different drug potencies, or can be derived from continuous activity models. Indeed, Swannstrom et al.

(2004) showed that a PSS based on continuous activity is superior to a score based on a cut-off.

Both the GSS and PSS suffer from ignoring drug-drug interactions. An alternative to these susceptibility scores has been suggested based on a very conservative model of drug-drug interactions (Beerenwinkel et al., 2003b). Whereas susceptibility scores assume that the effect of single drugs combines additively both within and across drug classes, the alternative model assumes that only the most active drug from each drug class determines the response to a drug combination. Using this model, the combination of ZDV, ddI and EFV in the aforementioned example (RT, K65R; protease, V32I and I47A; evaluated with ANRS) would be scored as  $1(\text{NRTIs})+1(\text{NNRTIs})=2$  because ddI is less effective than ZDV and thus ignored. At first sight, this approach implies no immediate benefit of having more than one drug from a class in a particular regimen as only the most active drug in each class is scored. However, in (Beerenwinkel et al., 2003b), not only is the current state of the viral population (as represented by the genotype) considered in scoring a drug combination, but also by traversing the evolutionary neighbourhood of the current genotype possible escape mutants. The most active drug from each class will usually vary across these escape mutants, which leads to 'indirect' positive interactions between compounds from the same class. These intraclass synergies are demonstrated using the example of a wild-type strain and the combination of ZDV/3TC, which differ significantly in their phenotypic resistance profiles. When taking into account only the current genotype (no consideration of escape mutants), the same score is assigned to the combination of both drugs as to the drugs in isolation. However, when the evolutionary neighbourhood is traversed and possible escape mutants are considered, the benefits of the combination can be observed. For example, consider the case of two escape mutants, one with a thymidine analogue mutation and the other with K65R. For the first mutant, 3TC will be an active drug, and for the second mutant, ZDV will be active. Thus, the combination of 3TC and ZDV will be active on both escape mutants, whereas the single-drug regimens would only be active on one of them.

As summarised above, scoring the mutational neighbourhood can improve response prediction. It seems intuitive that the neighbourhood search probes the genetic barrier to the drug combination and, as we have seen, the combination of ZDV/3TC has a higher genetic barrier than the individual drugs. However, the search, as it is performed by Beerenwinkel et al. (2003b), corresponds to a simple notion of genetic barrier, which is based on counting the number of substitutions necessary to acquire resistance (or lose activity). This simple definition ignores the different probabilities and timings at which these mutations occur. Thus, a more realistic assumption relies on defining the genetic barrier as the probability that the virus will not develop resistance within a given time period. An estimation method for this 'probabilistic' genetic barrier is described in (Beerenwinkel et al., 2005a). The application of genetic barriers in predicting response to combination therapy is described in (Beerenwinkel et al.,

2005b; Altmann et al., 2007), and implemented in the THEO (THERapy Optimizer) system<sup>2</sup>.

Another two-step approach for predicting response to combination therapy has been suggested based on the use of fuzzy logic (Prosperi et al., 2004), but there are also several approaches that try to relate baseline parameters directly with response to a drug combination. For example, the HIV Resistance Response Database Initiative is experimenting with neural networks, using mutations, drugs and viral load as direct inputs (Larder et al., 2004, 2005). Recently, another 'direct' approach consisting of a nearest neighbour classifier has been suggested (Prosperi et al., 2005): given baseline genotype, viral load and a drug combination, similar cases with known outcome are retrieved from a clinical database and the average of the outcomes is taken as the predicted response (as with the VircoType). The success of the method crucially depends on the similarity measure being used. Certainly, comparative evaluations are needed to identify the advantages and limitations of all these approaches.

### 5.1.7 Summary

This section has outlined the currently available interpretation algorithms for supporting the use of antiretroviral therapy. Several rule-based algorithms for predicting *in vivo* drug activity from the viral genotype are regularly updated by panels of experts and represent the most widely known tools to support decision-making. The label 'resistance interpretation algorithm' commonly assigned to these systems is misleading because all of them partially or completely rely on *in vivo* response data and should therefore be termed 'drug activity algorithms'.

Drug resistance, defined as the ability of the virus to replicate in the presence of drug, and measurable using phenotypic assays, is a major determinant of *in vivo* response. Consequently, approaches with and without cut-offs have been developed for predicting virological response from phenotypic resistance data. Although phenotypic resistance and viral replication capacity can be measured using experimental assays. These are more laborious and expensive than standard genotyping and thus less suitable for routine clinical use. Thus, it is desirable to be able to accurately predict both quantities from genotypic data and considerable advances have been made in this field.

Finally, in this chapter, the current approaches for predicting virological response to a drug combination have been reviewed. This is without doubt the 'holy grail' of the field, and we are still far from satisfying solutions. The mainstream approach is to use a susceptibility score in which the *in vivo* response to a drug combination is computed by adding the activities of the individual components. Alternatives have been proposed with the aim of modelling drug-drug interactions in non-additive ways, or of directly relating genotype to response with combination therapy, without resorting to individual

---

<sup>2</sup>[www.geno2pheno.org](http://www.geno2pheno.org)

drug activity models.

### 5.1.8 The future

Several topics are expected to play a crucial role as the field of interpretation algorithms moves forwards.

**Reference datasets and standardized evaluation schemes.** One important obstacle to progress in the field is the lack of large and publicly available clinical outcome data that can serve as a standard benchmark for the evaluation and comparison of interpretation algorithms. An additional problem is that the evaluation procedures differ between studies. As a consequence, it is difficult to compare reported performance measures (such as correlation coefficients or error rates) across different studies. Although comparisons of different methods are being published, they fall short of their objectives as long as they do not reach maximal reproducibility by publishing the underlying data, for example, via the Stanford HIV Drug Resistance Database<sup>3</sup>.

**Decision support for coreceptor antagonists (cf. Chapter 6).** The introduction of coreceptor inhibitors in clinical practice may require monitoring of coreceptor usage before, during and after treatment. Clinical practice would greatly benefit from reliable genotype-based prediction models for coreceptor usage, as these would be cheaper and faster than phenotypic assays. All genotypic approaches described to date are based on the third hypervariable (V3) loop of the envelope gp120, known as the major determinant of coreceptor usage. The most popular genotypic prediction is based on the classical 11/25 rule, predicting virus that uses the chemokine coreceptor CXCR4 (X4 virus; cf. Chapter 6 for details) in the presence of positively charged residues at positions 11 or 25 of the V3 region. However, several statistical learning methods have been shown to improve the sensitivity of detecting X4 virus (reviewed in [Jensen and Wout, 2003](#)). Recently, it was shown that the prediction performance can be further improved by incorporating immunological markers (such as CD4 counts) into predictive models as surrogate markers for undetected minority viral variants ([Sing et al., 2007b](#)), cf. also Chapter 6. Despite these advances, genotypic prediction of coreceptor usage needs further improvements, for example by considering positions outside the V3 region.

**Improved interpretability.** Paradoxically, many interpretation algorithms excluding rule-based algorithms lack easy interpretability (in the sense that a user can easily understand the rationale for a prediction), and are mainly used as ‘black box’ prediction

---

<sup>3</sup>[hivdb.stanford.edu](http://hivdb.stanford.edu)

systems. This is especially true with modern statistical learning methods such as neural networks and SVMs, although advances have been made in this respect (Sing et al., 2005b). To be clinically useful, the output of prediction systems should at least be augmented with well-calibrated confidence estimates. Moreover, sensitivity analyses and other techniques should be used to provide a certain degree of interpretability even for highly non-linear models such as neural networks.

**Additional baseline parameters for activity and response prediction.** Baseline genotype is the most popular piece of information for predicting activity, augmented by baseline viral load for predicting virological response. As more sophisticated models are being developed, the types of input that can be used by these models will expand. This might include (predicted) replication capacity, antiretroviral treatment history (Larder et al., 2004), adherence data (Larder et al., 2005) or immunological parameters. It remains a major challenge to integrate pharmacokinetic data, as obtained from therapeutic drug monitoring, into algorithms for predicting activity or response. The concept of inhibitory quotient, calculated as the ratio of the minimum plasma concentration ( $C_{min}$ ) for the drug to the level of drug resistance is a first step in this direction.

**Incorporating information on salvage options and therapy sequencing into regimen scoring.** Current algorithms predict virological response to a given drug combination. However, long-term planning requires considering the future treatment options should the regimen fail. Would there be enough remaining treatment options, or would further options be severely limited by the emergence of cross-resistance? Without doubt, questions related to optimal therapy sequencing will become increasingly important with the ever-rising number of approved anti-HIV-1 drugs and increasing number of individuals on long-term therapy.

#### Recommendations for clinical practice

- Interpretation algorithms are designed to support, not replace, the treating physician. The guidance they provide is further improved by expert opinion.
- Given the availability of several systems for different tasks, it is advisable not to rely on a single system, but to collect and consider recommendations from a variety of systems.
- Rule-based scores for predicting *in vivo* drug activity from genotype are currently the most widely used interpretation algorithms. However, as activity prediction is difficult, the output should be used in an informed way, along with other treatment-relevant information.

- Tools for predicting phenotype from genotype provide reliable predictions and can be used, in addition to other tools, to obtain a comprehensive overview of the current treatment situation.
- A variety of tools for predicting response to combination therapy are being developed (and one is already available), but these have not yet matured to the degree of reliability provided by systems for predicting phenotype from genotype.

## 5.2 Discovery and characterization of novel HIV drug resistance mutations<sup>4</sup>

In this section, we present a case study on the discovery of clinically relevant domain knowledge in the field of HIV drug resistance. Novel mutations in the HIV genome associated with treatment failure were identified by mining a relational clinical database. Hierarchical cluster analysis suggests that two of these mutations form a novel mutational complex, while all others are involved in known resistance-conferring evolutionary pathways. The clustering is shown to be highly stable in a bootstrap procedure. Multidimensional scaling in mutation space indicates that certain mutations can occur within multiple pathways. Feature ranking based on support vector machines and matched genotype-phenotype pairs comprehensively reproduces current domain knowledge. Moreover, it indicates a prominent role of novel mutations in determining phenotypic resistance and in resensitization effects. These effects may be exploited deliberately to reopen lost treatment options. Together, these findings provide valuable insight into the interpretation of genotypic resistance tests.

### 5.2.1 Introduction

**Motivation: Evidence for additional resistance-associated mutations and mutational clusters**

To date, the decision for follow-up drug combinations in patients failing therapy is routinely based on sequencing the relevant genomic region of the viral population harbored by the individual. The sequence is then analyzed to identify the presence of resistance-associated mutations for each of the 19 drugs currently available for anti-HIV therapy, by using mutation lists annually updated by the International AIDS Society (IAS) (Johnson et al., 2005) or other panels of human experts.

The situation is complicated by the fact that resistance mutations do not accumulate independently from each other. Rather, they are loosely time-ordered along mutational

---

<sup>4</sup>The work reported in this section was performed in collaboration with Valentina Svicher, Francesca Ceccherini-Silberstein, and Carlo-Federico Perno, University of Rome “Tor Vergata”.

pathways, leading to distinct mutational complexes or clusters.<sup>5</sup> Rational therapy planning is severely compromised by our limited understanding of these effects. Increasing evidence on additional mutations involved in the development of drug resistance (Gonzales et al., 2003; Svicher et al., 2005), besides those listed by the IAS, provides the incentive for our present study.

## Outline

We describe an approach towards the discovery and characterization of novel mutations associated with therapy failure from a large relational database, and their evolutionary and phenotypic characterization using supervised and unsupervised statistical learning methods. We focus on resistance against seven drugs from the class of nucleoside reverse transcriptase inhibitors (NRTIs), which target an HIV protein called reverse transcriptase (RT). This enzyme is responsible for translating the RNA genome of HIV back to DNA prior to its integration into the human genome. NRTIs are analogues of the natural building blocks of DNA, but lack a group essential for chain elongation. Thus, incorporation of a nucleoside analogue during DNA polymerization terminates the chain elongation process.

The knowledge discovery process described in this paper combines heterogeneous data from three different virological centers. To allow for integrated analysis, these data are stored in a relational database, whose structure is outlined in section 2. Systematic mining for mutations with differing propensities in NRTI-treated and untreated patients, respectively, as detailed in section 3, leads to the identification of 14 novel mutations associated with therapy failure. In section 4, we propose an approach towards characterizing the covariation structure of novel mutations and their association into complexes using hierarchical clustering and multidimensional scaling. Stability results are provided using a bootstrap method. Feature ranking based on support vector machines, described in section 5, allows for assessing the actual phenotypic impact of novel mutations. In section 6, we conclude by summarizing our approach, related work, and open problems.

### 5.2.2 Mining for novel mutations

Our approach towards identifying mutations associated with NRTI therapy is based on the assumption that these should occur with different frequencies in treatment-naive subjects and in patients failing therapy, respectively.

Thus, mining for novel mutations was based on contrasting the frequency of the wild-type residue with that of a specific mutation in 551 isolates from drug-naive patients and 1355 isolates from patients under therapy failure, at RT positions 1–320 (?). Chi-square tests were performed for all pairs of wild-type and mutant residues to de-

---

<sup>5</sup>Throughout this paper, the words *complex*, *cluster*, and *pathway* are used interchangeably.

termine mutations for which the null hypothesis that amino acid choice is independent from the patient population can be rejected. Correction for multiple testing was performed using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) at a false discovery rate of 0.05.

This procedure revealed 14 novel mutations significantly associated with NRTI treatment, in addition to those previously described in (Johnson et al., 2005): K43E/Q/N, E203D/K, H208Y, D218E were virtually absent in therapy-naives (< 0.5%), while K20R, V35M, T39A, K122E, and G196E were already present in the naive population with a frequency of > 2.5% but showed significant increase in treated patients. Surprisingly, mutations I50V and R83K showed significant decrease in the treated population as compared to therapy-naives.

### 5.2.3 Identifying mutational clusters

In this section we describe an unsupervised learning approach towards characterizing the covariation structure of a set of mutations and its application to the newly discovered mutations. Mutational complexes can give rise to distinct physical resistance mechanisms, but can also reflect different ways to achieve the same resistance mechanism. Indeed, the two most prominent complexes associated with NRTI resistance, the nucleoside analogue mutations (NAMs), groups 1 and 2, consisting of mutations M41L/L210W/T215Y and K70R/K219Q/D67N, respectively, both confer resistance via an identical mechanism, called primer unblocking. On the other hand, the multi-NRTI resistance complex with Q151M as the main mutation mediates a different physical mechanism in which recognition of chemically modified versions of the DNA building blocks is improved to avoid unintended integration. In essence, to appreciate the evolutionary role of novel mutations it is important to identify whether they aggregate with one of these complexes or whether they form novel clusters, possibly reflecting additional resistance mechanisms. This analysis was performed focusing on 1355 isolates from patients failing therapy.

#### Pairwise covariation patterns

Patterns of pairwise interactions among mutations associated with NRTI treatment were identified from the database using Fisher's exact test. Specifically, for each pair of mutations co-occurrence frequencies for mutated and corresponding wild-type residues were contrasted in a 2-way contingency table, from which the test statistic was computed.

A visual summary of these pairwise comparisons, part of which is shown in Fig. 5.1, immediately reveals the classical mutational clusters described above. It is also apparent that no significant interactions are formed between the Q151M complex and mutations from the NAM clusters, suggesting that resistance evolution along the for-

mer pathway is largely independent from the other complexes and that different pathways may act simultaneously on a sequence, at least if they mediate different physical resistance mechanisms.

In contrast, significant interactions take place across the two NAM complexes. Antagonistic interactions between the core NAM 1 mutations L210W / M41L / T215Y and NAM 2 mutations K70R and K219Q might indicate negative effects of simultaneous evolution along these two pathways, which both contribute to the primer unblocking mechanism.

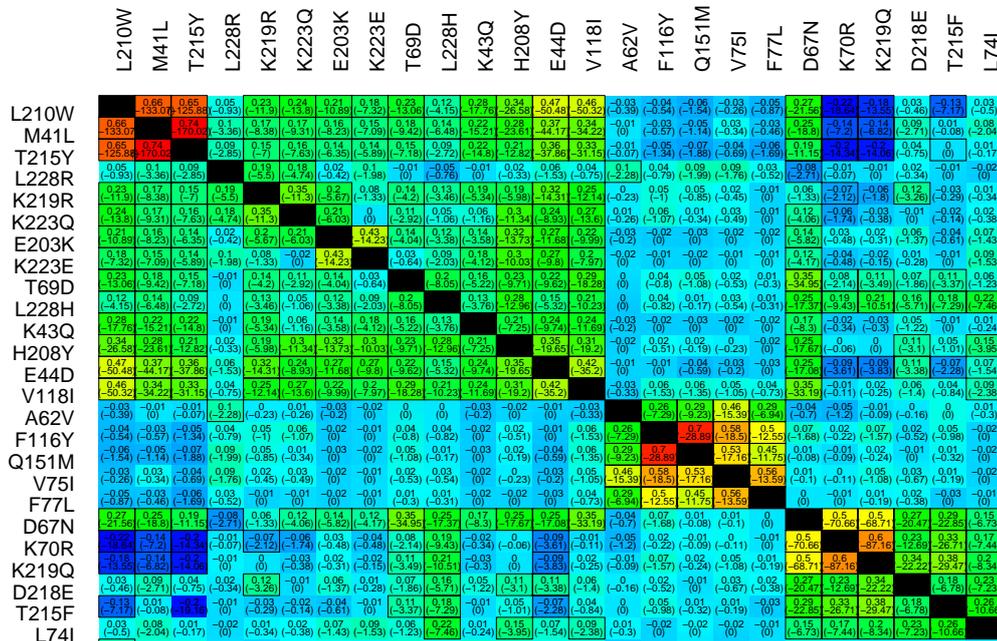


Figure 5.1: Pairwise  $\phi$  correlation coefficients between mutations (part view), with red indicating maximal observed positive covariation and blue maximal observed negative covariation. Boxes indicate pairs whose covariation behavior deviates significantly from the independence assumption, according to Fisher’s exact test and correction for multiple testing using the Benjamini-Hochberg method at a false discovery rate of 0.01. The classical mutational complexes introduced in section 4 form distinct clusters, from left to right: NAM 1, Q151M multi-NRTI, NAM 2.

Clustering mutations

Dendrograms obtained from hierarchical clustering allow for a more detailed analysis of mutation covariation structure. The similarity between pairs of mutations was assessed using the  $\phi$  (Matthews) correlation coefficient, as a measure of association between two binary random variables, with 1 and  $-1$  representing maximal positive

and negative association, respectively. This similarity measure was transformed into a dissimilarity  $\delta$  by mapping  $\phi = 1$  to  $\delta = 0$  and  $\phi = -1$  to  $\delta = 1$ , with linear interpolation in between. Since it is impossible to obtain adequate dissimilarity estimates for pairs of mutations at a single position from cross-sectional data,<sup>6</sup> these were treated as missing values in our approach. The resulting partial dissimilarity matrix was taken as the basis for average linkage hierarchical agglomerative clustering.<sup>7</sup>

The dendrogram in Fig. 5.2 reveals that most novel mutations group within the NAM 1 cluster (T215Y/M41L/L210W), except for D218E and F214L, which aggregate to NAM 2. Interestingly, mutations R83K and I50V, which occur more frequently in naive than in treated patients appear to form a novel outgroup.

To assess the stability of the dendrogram, 100 bootstrapped samples of RT sequences were drawn from the original 1355 sequences. Distance calculation and hierarchical clustering were performed for each of these samples as described above. Then, for each subtree of the dendrogram in Fig. 5.2, the fraction of bootstrap runs was counted in which the *set* of mutations defined by the subtree occurred as a subtree, without additional mutations.<sup>8</sup>

The four edge weights next to the root of the dendrogram show that the reported association of mutations D218E and F214L with NAM 2 is indeed highly stable across resampled data subsets, as is the grouping of other novel mutations with NAM 1, and the outgroup status of R83K and I50V. Bootstrap values for the lower dendrogram levels have been omitted for the sake of clarity; they range from 0.35 to 0.99, reflecting considerable variability of intra-cluster accumulation order. Finally, the core NAM 1 and NAM 2 mutations, respectively, are again grouped together with maximal confidence.

### Multidimensional scaling in mutation space

As can be seen in Fig. 5.1, certain mutations correlate positively with mutations from both NAM pathways – an effect which might be missed in a dendrogram representation, and which can be visualized, at least to some extent, using multidimensional scaling (MDS).

The goal in MDS is, given a distance matrix  $D$  between entities, to find an embedding of these entities in  $\mathbb{R}^n$  (here  $n = 2$ ), such that the distances  $D'$  induced by the embedding match those provided in the matrix optimally, defined via minimizing a particular “stress” function. Our embedding is based on the Sammon stress function

---

<sup>6</sup>Such mutation pairs never co-occur in a sequence.

<sup>7</sup>In average linkage with missing values, the distance between clusters is simply the average of the *defined* distances.

<sup>8</sup>Thus, in computing confidence values increasingly closer to the root, topology of included subtrees is deliberately ignored (otherwise, values would be monotonically decreasing from leaves to the root).

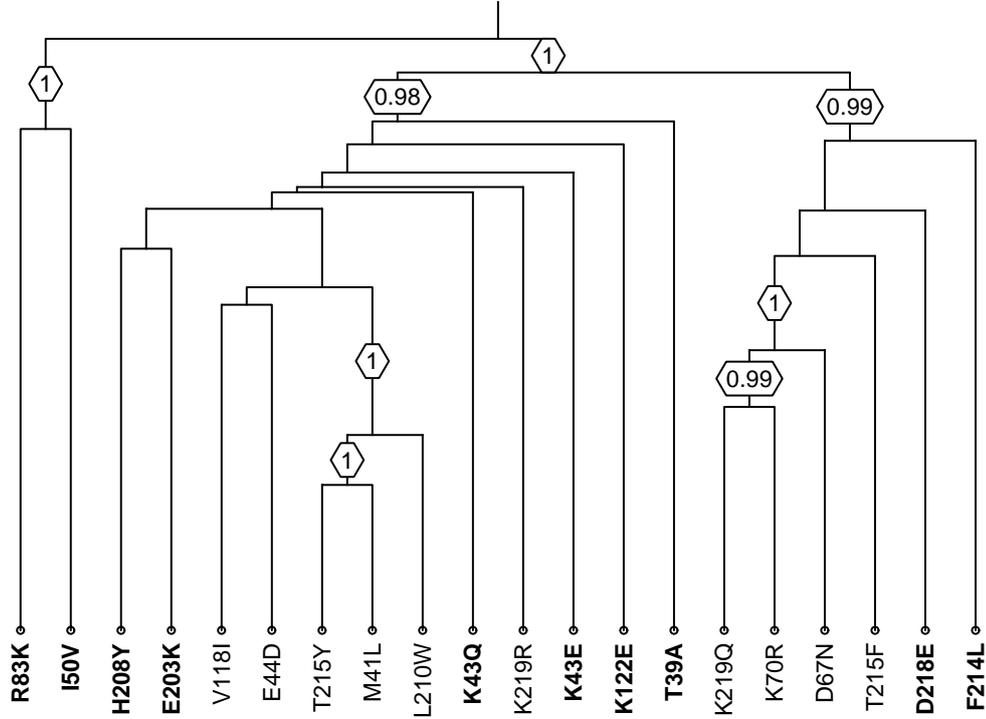


Figure 5.2: Dendrogram, as obtained from average linkage hierarchical clustering, showing the clear propensity of novel mutations to cluster within one of the classical NAM complexes T215Y/M41L/L210W and K219Q/K70R/D67N, or in the case of R83K and I50V, to a distinct outgroup. Novel mutations are marked in bold face. Bootstrap values which are not relevant for our discussion have been removed for the sake of clarity. Distances between mutations at a single position are treated as missing values in the clustering procedure. Remarkably, such pairs of mutations can show differential clustering behavior, as is apparent in the case of K219Q/R and T215F/Y.

(Sammon, 1969),

$$E(D, D') = \frac{1}{\sum_{i \neq j} D_{ij}} \sum_{i \neq j} \frac{(D_{ij} - D'_{ij})^2}{D_{ij}}, \quad (5.1)$$

which puts emphasis on reproducing small distances accurately. As in clustering, mutation pairs at a single position are excluded from the computation of the stress function, to avoid undue distortions.

The optimal Sammon embedding for the mutation distance matrix derived from pairwise  $\phi$  values is shown in Fig. 5.3. Note that due to the non-metricity of this matrix, which violates the triangle inequality, such an embedding cannot be expected

to preserve all original distances accurately. Still, the MDS plot supports the main conclusions from Section 5.2.3, such as to the structure of the classical NAM complexes, the outgroup status of R83K and I50V, and the exclusive propensity of certain mutations, such as K43E/Q or F214L, to a unique pathway. In addition, the plot also suggests a role in both NAM pathways for several mutations, such as H208Y, D67N, or K20R.

## 5.2.4 Phenotypic characterization of novel mutations using SVM-based feature ranking

The analyses described above allowed us to associate novel mutations with treatment failure and to group them into distinct mutational complexes. In this section we address the question whether novel mutations contribute directly to increased resistance or merely exert compensatory functions in removing catalytic deficiencies induced by the main resistance-conferring mutations. We do so by analyzing their role in classification models for predicting phenotypic drug resistance.

Resistance of a given HIV strain against a certain drug can be measured *in vitro* by comparing the replicative capacity of the mutant strain with that of a non-resistant reference strain, at increasing drug concentrations (Walter et al., 1999). The result of this comparison is summarized in a scalar *resistance factor*. On the basis of 650 matched genotype-phenotype pairs for each drug, we have built predictive models, using decision trees (Beerenwinkel et al., 2002), and support vector machine classification and regression. These models are implemented in a publically available web server called *geno2pheno*<sup>9</sup> (Beerenwinkel et al., 2003a), which has been used over 36,000 times since December 2000<sup>10</sup>.

While support vector machines are widely considered as the state-of-the-art in prediction performance, there is a common attitude that these models are difficult to interpret and suffer from “the same disadvantage as neural networks, viz. that they yield black-box models” (Lucas, 2004). However, a substantial set of techniques is in fact available for feature ranking with SVMs (e.g. (Guyon et al., 2002)), by removing features or destroying their information through permutation, and even for extracting rule sets from SVMs.

In our case, using the linear kernel  $k(x, y) = \langle x, y \rangle$  (standard nonlinear kernels did not significantly improve accuracy), feature ranking is particularly straightforward. Due to the bilinearity of the scalar product, the SVM decision function can be written

---

<sup>9</sup><http://www.geno2pheno.org>

<sup>10</sup>This server is updated regularly and is now based on a larger number of genotype-phenotype samples per drug than at the time at which this study was performed.

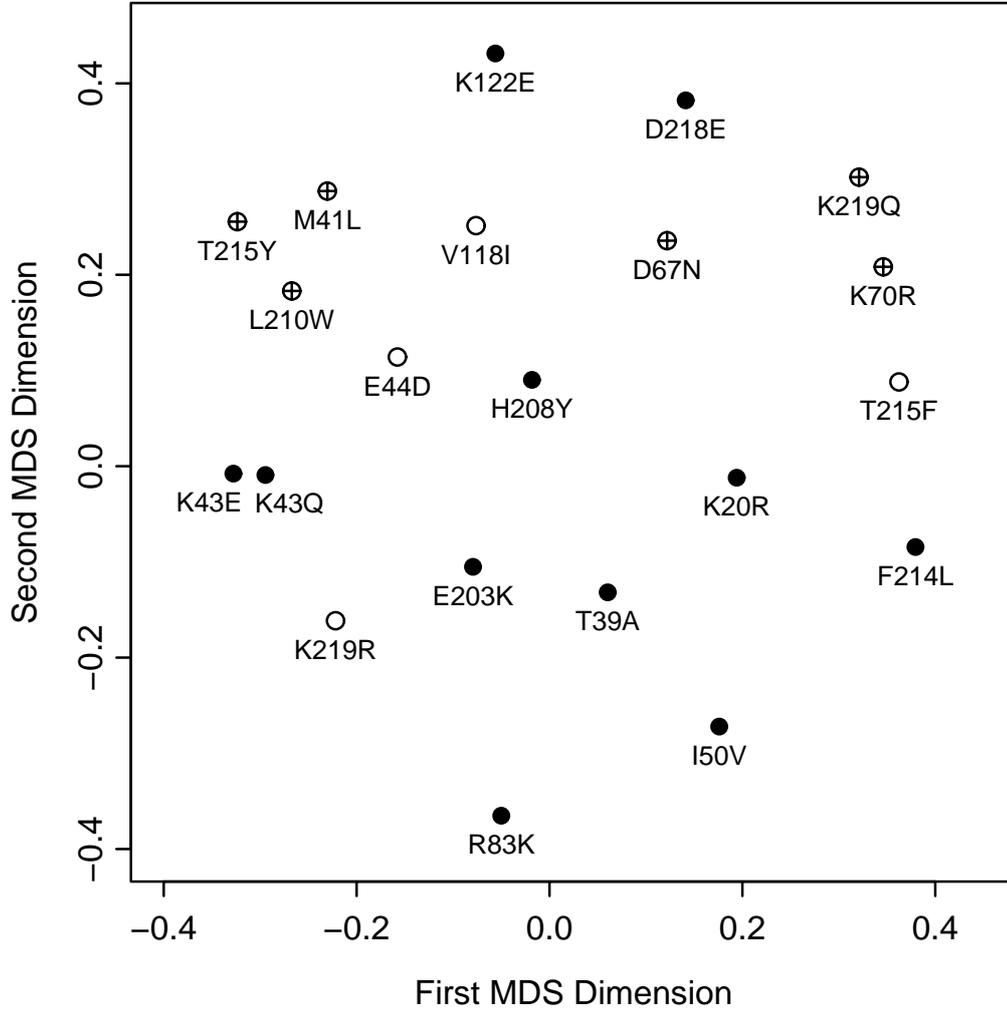


Figure 5.3: Multidimensional scaling plot of novel (shown in black) and classical mutations (in white; main NAMs indicated by a cross), showing a two-dimensional embedding which optimally (according to Sammon's stress function) preserves the distances among the mutations, as derived from the  $\phi$  correlation coefficient. Distances between mutations at a single position were treated as missing values.

as a linear model,

$$f(x) = \sum_i y_i \alpha_i k(x_i, x) + b = \langle \sum_i y_i \alpha_i x_i, x \rangle + b, \quad (5.2)$$

allowing for direct assessment of the model weights.

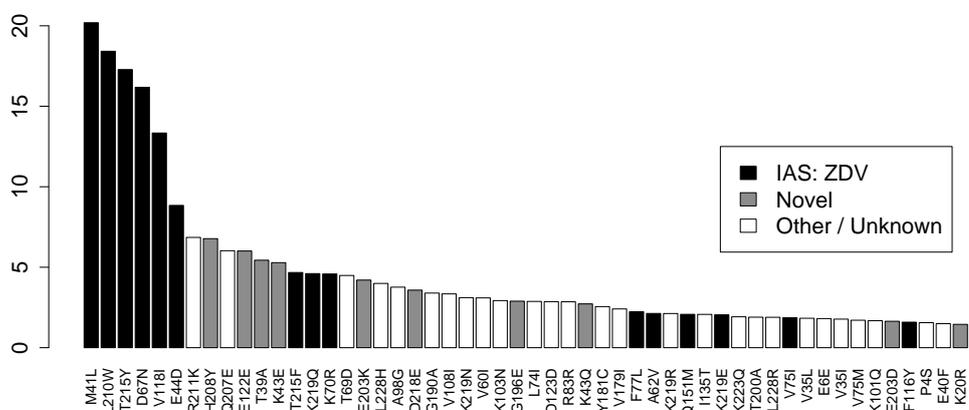


Figure 5.4: Major mutations conferring resistance to zidovudine (ZDV), as obtained from SVM-based ranking of 5001 mutations. Bar heights indicate z-score-normalized feature weights (for example, mutation M41L is more than 20 standard deviations above the mean feature weight). Mutations associated with ZDV resistance by the International AIDS Society are shown in black; novel mutations identified from frequency comparisons in treated and untreated patients are shown in grey. Mutations not further studied here are shown in white.

Figure 5.4 shows the result of this SVM-based feature ranking for zidovudine (ZDV), one of the seven NRTIs. All mutations associated with resistance to ZDV in the current resistance update provided by the International AIDS Society (Johnson et al., 2005) appear in the top 50 of 5001 features (250 positions, 20 amino acids each, plus 1 indicator for an insertion), with the first six positions exclusively occupied by classical NAM mutations (shown in black). This observation provides evidence that our models have adequately captured established domain knowledge as contributed by human experts. Remarkably, when investigating the role of novel mutations (shown in grey) in the model, we find that many of them are prominently involved in determining ZDV resistance, ranking even before several of the classical ZDV mutations.

These findings generalize to the whole NRTI drug class, as is obvious from Table 5.1, which shows the ranks of novel mutations in the individual drug models. Table 5.1 also reveals some striking and unexpected differences among mutations. For example, various results suggest a close relationship of mutations H208Y and E203K, which form a tight cluster in the dendrogram, show up as neighbors in the multidimensional scaling plot, and exhibit similar rank profiles – with the notable exception of their differential impact on ddC resistance.

This surprising difference and other effects are more readily appreciated in Fig. 5.5, which shows the weights associated with novel mutations in the individual SVM drug models (after drug-wise z-score weight normalization for improved comparabil-

	ZDV	ddI	ddC	d4T	3TC	ABC	TDF
R83K	4972	3722	718	79	4973	539	154
I50V	4910	803	4702	4855	4736	4818	4899
H208Y	8	16	170	9	114	20	65
E203K	17	271	4963	103	8	19	103
K43Q	30	121	72	684	19	32	18
K43E	12	19	641	10	107	49	10
K122E	10	21	37	45	72	72	774
T39A	11	3814	4882	528	169	4017	50
D218E	20	22	103	50	25	13	659
F214L	119	898	4019	735	128	303	4844
M184V	67	2	1	4971	1	1	4994

Table 5.1: Ranks of novel mutations in SVM models for seven NRTIs, with rank 1 indicating maximal contribution to resistance, and rank 5001 maximal contribution to susceptibility. The classical mutation M184V is shown here for comparison, due to its particularly strong resensitization effect. The clinical (but not virological) relevance of results concerning ddC is limited by the limited popularity of this drug.

ity). Indeed, increased resistance against ZDV, 3TC, and ABC upon appearance of E203K seems to coincide with *resensitization* (i.e. increased susceptibility) towards ddC. A similar, even more extreme effect can be observed in the case of T39A, for which increased resistance against ZDV and TDF again contrasts with increased ddC susceptibility. R83K shows dual behavior: increased d4T resistance and increased ZDV susceptibility. The presence of I50V is associated with increased susceptibility against all NRTIs, explaining its decreased frequency in treated patients.

Related effects have attracted considerable recent interest due to their possible benefits in reopening lost treatment options (Wang et al., 2004b). Arguably the most pronounced behavior can be seen in the classical mutation M184V (Table 5.1), known to confer high-level resistance to 3TC but inducing d4T and TDF resensitization. SVM-based feature ranking reproduces this effect in a most striking manner: For ddI, ddC, 3TC, and ABC, M184V turns out to be the top resistance mutation, with contributions of 11.2, 15.4, 42.0, and 20.8 standard deviations above the mean. In contrast, the same mutation appears to be one of the major contributors of increased susceptibility towards d4T and TDF, 3.5 and 8.2 standard deviations *below* the mean, respectively.

### 5.2.5 Discussion

We have presented a case study on mining a multi-center HIV database using supervised and unsupervised methods. Previously undescribed mutations could be asso-

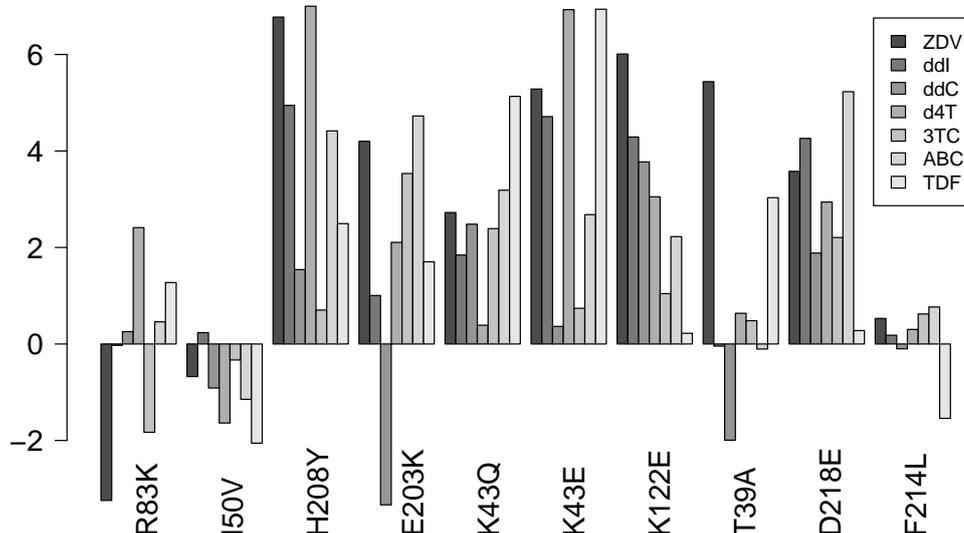


Figure 5.5: Weights of novel mutations (after z-score normalization) in SVM models for seven NRTIs. For example, mutation E203K contributes significantly to ZDV resistance, while increasing susceptibility towards ddC.

ciated with resistance towards the drug class of nucleoside reverse transcriptase inhibitors and grouped into mutational clusters. SVM-based feature ranking on an independent data set suggests a direct contribution of novel mutations to phenotypic resistance and an involvement in resensitization effects which might be exploited in the design of antiretroviral combination therapies. Feature ranking has been incorporated into the geno2pheno system [Figure 5.6](#).

**Mutation screening.** Novel mutations were found by position-wise comparisons, leaving inter-residue effects aside. It is conceivable that additional sets of mutations related to therapy failure, whose effect is too weak to discern in isolation, could be identified using other methods, such as discriminating item set miners. In fact, we have recently proposed an approach towards mining discriminating item sets, in which an overall rule weight in a mixture model of rules is modulated by the genomic background in which a rule matches ([Sing et al., 2004](#)). Further work will have to explore the possible benefits of using such strategies in the present context.

**Covariation versus evolution.** Dendrograms and MDS analyses describe the association of mutations into mutational complexes, but refrain from explicit statements on the accumulation order of mutations. Other approaches, most notably mutagenetic tree models ([Beerenwinkel et al., 2005a](#)), are explicitly tailored towards elucidating HIV evolutionary pathways from cross-sectional data as those used in our study. However, while novel mutations exhibit distinct clustering behavior, the actual order of

## III. Phenotype prediction

Drug	Resistance Factor RF (*)	z-score	Scored Mutations (**)
ZDV	377.4	13.0	215Y 70R 41L 62V 67N 197H 35T 49N 39S
ddl	3.6	6.5	75T 62V 215Y 162C 214F 135V 58S 179I 207E 216P 163T 121E 49N 4L 69D 60I
d4T	2.8	6.1	67N 215Y 162C 62V 41L 69D 122K 216S 39S 207E 35T 159M
3TC	9.6	7.2	215Y 41L 162C 207E 108I 62V 70R 67N 35T 121E 58N 60I 179I 163T 216P 2F 181F
ABC	4.6	9.5	215Y 41L 62V 35T 179I 67N 58N 162C 214F 69D 70R 135V 108I 39S 87L 4H
TDF	7.8	12.2	215Y 62V 35T 41L 70R 67N 135V 58I 207E 60I
NVP	10.6	3.3	135V 210F 197H 4L 35T 108I 39A 75T 121E 163I 62V 181F 58S 122K 83S 2L 60I 179I
EFV	4.4	3.2	135V 62V 108I 58I 35T 39A 181F 4L 75T 69D 210F 70R 87L 207E 214F 41L 60I 122K 163N 59Q 179I
SQV	33.5	12.7	90M 73S 2K 6G 72V 43T 71T 5V 23R 59F 63P 37N
IDV	10.9	6.8	90M 63P 73S 71T 20I 4I 72V 62V 43T
RTV	30.3	11.4	90M 43T 33F 73S 4I 2K 23R 63P 61E
NFV	31.2	7.5	90M 63P 20I 72V 43T 33F 2K 71T 73S 4I 3I
FPV	5.1	4.0	33F 90M 63P 2K 43T 61E 72V
LPV	4.7	4.9	33F 63P 90M 71T 20I 73S 62V 72V
ATV	12.3	6.6	90M 73S 33F 36I 63P 20I

Figure 5.6: SVM-based feature ranking in the PDF output of the geno2pheno system. The column “Scored Mutations” shows those mutations with the highest absolute SVM weights that contribute either positively (shown in red) or negatively (green) to predicted resistance levels. Mutations are shown up to 95% of the total mutation weight mass.

their accumulation seems to be relatively flexible, challenging the applicability of such evolutionary models in this setting.

**SVM-based versus correlation-based feature ranking.** To date, feature ranking is performed mostly using simple correlation methods, in which features are assessed in their performance to discriminate between classes *individually*, e.g. by using *mutual information*. However, as detailed in (Guyon et al., 2002), feature ranking with correlation methods suffers from the implicit orthogonality assumptions that are made, in that feature weights are computed from information on a single feature in isolation, without taking into account mutual information between features. In contrast, statistical learning models such as support vector machines are inherently multivariate. Thus, their feature ranking is much less prone to be misguided by inter-feature dependencies than simple correlation methods. Further analysis of the feature rankings induced by different methods can provide valuable insights into their particular strengths and weaknesses and suggest novel strategies for combining models from different model classes.

## 5.3 Inferring response from genotype: with or without predicted phenotypes?<sup>11</sup>

In the prediction of response to antiviral combination therapy, we compare the use of predicted phenotypes as an “intermediate” step with approaches based on a direct correlation of genotype with response. In contrast to other reports, we find that the use of predicted phenotypes is very competitive, often outperforming direct approaches. We show that synergies between phenotype-based and direct representations can be used to build “hybrid” systems combining both sources of information.

### 5.3.1 Introduction

Modern antiretroviral combination therapy can substantially delay disease progression, prolong survival and maintain quality of life, but a cure for HIV infection remains out of reach. Therefore, research focuses not only on the search for novel drugs, but also on exploiting the currently available drug collection to the best possible effect using personalized therapy administration. The main obstacle to ultimate treatment success is the ability of the virus to rapidly acquire mutations that confer resistance to specific drugs.

Prior to the availability of assays for determining drug resistance, drug combinations had to be chosen exclusively based on the clinical and therapeutic history of the patient. Nowadays, in many countries, the genomic make-up of a viral population at treatment failure (and increasingly also in newly diagnosed patients) is routinely determined by sequencing the relevant portions of the HIV genome. The resulting data are often collected in regional or national databases and have led to the discovery of an ever-increasing number of resistance-associated mutations, which tend to occur in diverse mutational patterns. The interpretation of a given pattern with respect to its implications on drug resistance and response to combination therapy is an extremely complex task. Consequently, a number of expert panels have been developing and continue to refine rules- or mutation score-based interpretation algorithms. Backed up by several resistance and therapeutic clinical trials (reviewed in [Descamps and Brun-Vezinet, 2006](#)), the benefit of using viral genomic information in addition to clinical and therapeutic data, as opposed to exclusively relying on the latter, is no longer controversial. However, the diversity and complexity of resistance-associated mutational patterns and their impact on the activity of drugs given in combination is far from understood.

While sequencing assays are the basic tool for detecting mutations in the viral

---

<sup>11</sup>The work reported in this section was performed in collaboration with Virco/Tibotec, Andre Altmann (MPI for Informatics, Saarbrücken), and using clinical data kindly provided by Bob Shafer (Stanford).

population, they do not provide any information about the clinical role of these mutations. A standard approach for distinguishing resistance-associated mutations from mere polymorphisms has been to compare mutation frequencies in untreated versus treated patients (reviewed in [Sabin, 2006](#)). However, this approach cannot provide quantitative assessments of the phenotypic effect of particular mutations and is thus of limited utility in the development of prediction algorithms. In order to develop quantitative models it is necessary to link genotypic and treatment information to virological response data reflecting the activity of a given regimen on a viral population with a given genomic make-up. Such a link between a viral genotype, a drug combination, and a measure of how effectively this particular drug combination works for this particular genotype, is the atomic piece of information in datasets used for building models for virological outcome prediction. Adopting a terminology coined by the HIV Resistance Response Database Initiative (RDI), we shall call this basic data unit a treatment-change episode (TCE). It is well known that other pieces of clinical and therapeutic information, such as baseline viral load or CD4<sup>+</sup> T-cell counts, treatment and adherence history, can contribute to an improved prediction. However, as this information is not always available, models should always be built in such a way that therapies can be ranked based on the minimum of only a viral genotype.

TCE datasets extracted from clinical databases can and have been used for model building without any further processing. Expert panels screen these data for mutations associated with treatment failure and can then integrate this information together with other sources of knowledge, including scientific publications, into their carefully hand-crafted interpretation tables. As an alternative approach, an increasing number of research groups are trying to develop well-defined, algorithmic and strictly reproducible approaches to model building based exclusively on the foundation of statistics, statistical learning, and bioinformatics. While our own work follows this latter paradigm, we emphasize that none of the approaches to model building is inherently superior and each has its advantages and disadvantages (which we shall outline in [Section 5.3.4](#)). Thus, in evaluating and comparing the clinical utility of different approaches, we are neither able nor willing to assess a particular design philosophy (i.e. human-based versus automated). Rather, the only quantity of interest in this paper is the predictive performance, as quantified by a variety of standard performance measures.

While TCEs can be used "as is" for model building, as described in the previous paragraph, it should be evaluated whether additional sources of data can be used during model building to improve the genotype-based prediction models. Importantly, the emphasis is on using the additional data only during model building, as opposed to using them also during prediction, as would be the case in the example of using baseline viral load or treatment history as additional predictors. The additional data are only used to extract all possible information from the genotype during model building. As a consequence, the final models still remain useful when only a genotype is available. Examples of this approach related to HIV include the use of x-ray protein

structure data in combination with molecular modeling to improve to prediction of phenotypic drug resistance (Draghici and Potter, 2003; Jenwitheesuk and Samudrala, 2005; Shenderovich et al., 2003) or coreceptor usage (Sander et al., 2007), or the use of evolutionary modeling via the "genetic barrier" to resistance to improve the prediction of response to combination therapy (Altmann et al., 2007).

In this study, we focus on the use of a particular source of additional information during model building that has recently caused an unexpected controversy: phenotypic resistance assays. Intuitively, the output of these assays is immediately appealing, as compared to genotype-clinical databases. First, different from the clinical setting, they provide information about the resistance of virus against individual drugs. Second, the fold-change in 50% inhibitory concentration, as compared to a susceptible wild-type strain, provides a much simpler measure of resistance than a complex mutational pattern. As experimental phenotyping is too expensive for routine clinical use, the use case is exactly as described in the previous paragraph: to build genotype-phenotype models which can then be used on a given genotype as an additional source of information in predicting virological outcome. However, three factors have contributed to general doubts about the usefulness of phenotypes in the clinical setting: First, the lack of consistent improvements of phenotype-based treatment decisions as compared to the use of genotypic resistance testing or even to the previous standard of care (clinical and therapeutic history) in a number of clinical trials (reviewed in Descamps and Brun-Vézinet, 2006). Second, the insight that mutations in the genotype might not have any phenotypic effect, while still providing information about the extent of previous resistance or evolutionary progression, such as silent mutations on the nucleotide level, revertants, or polymorphisms associated with specific resistance pathways. Third, the observation that some mutations might have a different effect in the *in vitro* resistance assays than *in vivo*. Taken together, these factors have led to a frequently uttered mainstream view in the drug resistance community that "attempting to infer response from genotype via the intermediate step of predicting the phenotype is likely to have serious limitations" (Larder et al., 2007; Brun-Vézinet et al., 2004). In this chapter, we re-examine and challenge this view in a detailed evaluation of the relative merits and weaknesses of genotype- and phenotype-based approaches in virological response prediction. In particular, we address the two questions: (1) How is the performance of approaches exclusively based on predicted phenotypes in comparison to established genotypic interpretation algorithms? (2) What is the potential for improvements by hybrid approaches using both the genotypic-clinical, as well as the genotype-phenotype information?

## 5.3.2 Materials and Methods

### Standard datum definitions

The virological response to a therapy was dichotomized into success and failure. For the present analysis, which focuses on the intrinsic value of alternative input representations, this allows us to compare similarities and differences between the various representations in the most intuitive way. In this study, our main focus is on a genotype-centric standard datum defined by the EuResist consortium (Zazzi, 2006). This definition is as close as possible to the notion of “clinical”, as opposed to “phenotypic” drug resistance (cf. Section 5.3.4). A diagrammatic representation is shown in Figure 5.7. Any available genotype is considered as evidence of a failing regimen, because, in general, sequencing can only be performed if the virus load exceeds  $\approx 1,000$  copies per ml. Intuitively, this means that the drug combination at the time of sequencing must be considered a bad therapy for this particular genotype (otherwise sequencing would not have been possible). Successful regimens are defined by inspecting therapies that follow a failure: If the virus is undetectable at least once during the course of the follow-up therapy and if the most recent sequencing was performed no earlier than three months before starting the therapy, then the respective treatment was considered a success. If multiple genotypes are available, the most recent sequence sample before the onset of therapy was used. While this definition is most appropriate for the question at hand, we also performed all experiments in parallel for the very different “classical” standard datum definition of the EuResist project (Zazzi, 2006). Briefly, in the classical definition, success or failure is defined based on the viral load measurement that is closest to 56 days after onset of therapy, among those available within 28 to 84 days after onset. If that value is below 500 copies per ml, the therapy is considered a success, otherwise a failure.

### Data

With the same motivation as that for using two TCE definitions, we perform our analyses in parallel on two disjoint databases. The first dataset, called “Stanford” henceforth, consists of data from two Northern California clinic populations undergoing genotypic resistance testing at Stanford University and of data from the Stanford HIV Drug Resistance Database (clinical studies ACTG 320, ACTG 364, GART, and HAVANA). The second dataset, called “EUResist” consists of the EUResist database, which merges routine clinical data from the German database Arevir, the Italian database ARCA, and the Swedish database maintained at Karolinska University Hospital. The Stanford raw data consists of 25,717 therapies; 16,288 sequences; 6,706 patients; 110,392 viral load measurements; giving rise to 6,337 TCEs (4,776 failures; 1,583 successes) according to the genotype-centric, and 2,351 TCEs (924 failures; 1,427 successes) according to the standard TCE definition. The EUResist data set contains 34,078 ther-

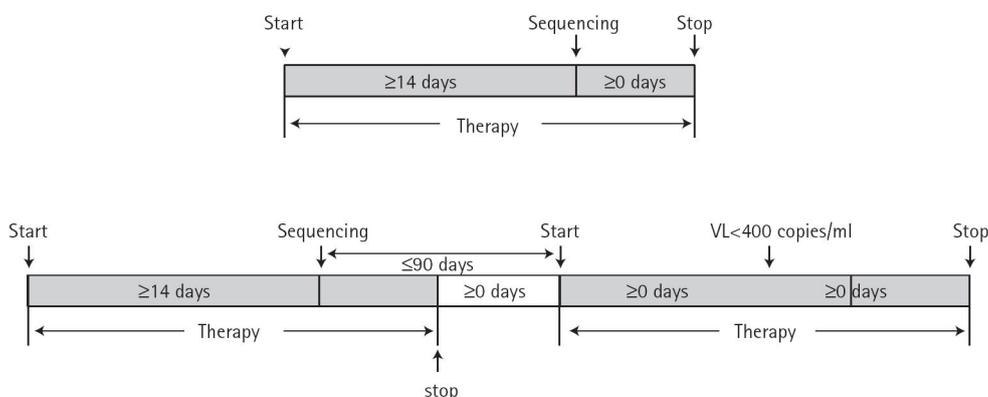


Figure 5.7: The “genotype-centric” EuResist standard datum. A treatment change episode (TCE) resulting in sequencing is always considered to be a failure (top part of the figure). If the virus is undetectable during the treatment following a failure, the TCE is called a success (bottom part of the figure).

opies; 13,628 sequences; 17,162 patients; 166,037 viral load measurements; giving rise to 5,224 TCEs (4,320 failures; 904 successes) according to the genotype-centric, and 1,064 TCEs (450 failures; 614 successes) according to the standard TCE definition. We focus on therapies consisting exclusively of the following 17 drugs from the different classes: nucleoside reverse transcriptase inhibitors (NRTIs): ZDV, ddI, ddC, d4T, 3TC, ABC, TDF; non-nucleoside reverse transcriptase inhibitors (NNRTIs): NVP, DLV, EFV; protease inhibitors (PIs): SQV, RTV, IDV, NFV, APV, LPV, ATV.

## Inputs

The objective of this study is to compare the reliability of predicting success or failure of antiretroviral combination therapy when using various alternative input representations. Since the use of predicted phenotypes has been criticized frequently, our particular attention is on comparing this input representation with other representations, and to investigate potential synergies by hybrid approaches. In particular, we investigate nine input representation: three expert algorithms, one representation relying on predicted phenotypes, one relying on raw genotypes, and four “hybrid” approaches combining either expert algorithms or raw genotype with phenotype:

1. **ANRS.** One number for each drug, representing the predicted activity of the drug for a given genotype. Activity was scaled between 0 and 1, with 0 indicating fully inactive drugs and 1 fully active drugs. Prediction was performed using the ANRS expert algorithm, version 2006/07 (Section 5.1.3). Mapping of the categorical ANRS predictions to activity levels was as follows: “Resistance”  $\mapsto$  0; “Possible resistance”  $\mapsto$  0.5; “Susceptible”  $\mapsto$  1. All drugs *not* applied

in a given treatment change episode were scored as 0 (i.e., fully inactive).

2. **Rega.** Same as (1), but using the Rega expert algorithm, version 6.4.1 (Section 5.1.3). Mapping of Rega resistance levels to drug activity levels in the interval  $[0, 1]$  was performed as follows: “Resistant”  $\mapsto 0$ ; “Intermediate resistant”  $\mapsto 0.5$ ; “Susceptible”  $\mapsto 1$ .
3. **HIVDB.** Same as (1), but using the Stanford HIVDB expert algorithm. As we introduced in Section 5.1.3, the HIVDB algorithm is different from other expert algorithms in that it does not assign categorical resistance levels directly. Rather, mutations in the genotype are assigned individual scores, which are then added into an overall genotype score (in this way, HIVDB can be regarded as an expert-derived linear regression model). As part of HIVDB, these scores are then discretized into five levels of resistance. The mapping of these resistance levels to drug activity levels was as follows: “High-level resistance”  $\mapsto 0$ ; “Intermediate resistance”  $\mapsto 0.25$ ; “Low-level resistance”  $\mapsto 0.5$ ; “Potential low-level resistance”  $\mapsto 0.75$ ; “Susceptible”  $\mapsto 1$ .
4. **Pheno.** Same as (1), but using the VircoTYPE 4.0 system for predicting phenotypic resistance to the individual drugs for a given genotype. The VircoTYPE is a linear model with pair-wise interaction terms that has been fitted to an average of 46,100 matched genotype-phenotype pairs per drug. For a given genotype, the prediction is  $\log_{10}(\text{IC}_{50})$ . This “resistance factor” was then mapped into the interval  $[0, 1]$  as follows. For each drug, VircoType provides an upper and a lower cutoff. Resistance predictions above the upper cutoff were mapped to a drug activity level of 0 (fully inactive drug), while resistance predictions below the lower cutoff were mapped to a drug activity level of 1 (fully active drug). Resistance predictions between these cutoffs were transformed into activity levels in the range  $[0, 1]$  by simple linear interpolation ( $\text{RF} \mapsto 1 - (\text{RF} - \tau_L) / (\tau_U - \tau_L)$ ), with  $\tau_U$  and  $\tau_L$  denoting the upper and lower cutoff, respectively. Like with the three expert algorithms, drugs not applied in a given treatment change episode were assigned an activity level of 0.
5. **Geno.** This is a 0/1 representation of the genotype that indicates the presence (“1”) or absence (“0”) of specific mutations. We used the list of mutations suggested by the International AIDS Society (Fall 2006 version, (Johnson et al., 2006)). Accumulating the mutations listed for the 17 drugs considered here, this resulted in a total of 62 protease and 32 reverse transcriptase mutations. In the expert and Pheno representations, the drugs applied in a regimen were encoded implicitly, by setting predictions for non-occurring drugs to fully inactive. In the Geno representation, however, we need an additional 17 dimensions to indicate the presence (“1”) or absence (“0”) of individual drugs. Thus, in total a genotype

plus drug combination is represented by a  $(62 + 32 + 17)$ -dimensional binary vector.

6. **ANRS+Pheno.** This is a “hybrid” input representation concatenating the ANRS and Pheno representations introduced above, leading to a  $(17 + 17)$ -dimensional vector.
7. **Rega+Pheno.** Concatenation of the Rega and Pheno representation.
8. **HIVDB+Pheno.** Concatenation of the HIVDB and Pheno representation.
9. **Geno+Pheno.** Concatenation of the Geno and Pheno representation. As the Pheno part captures the drug combination implicitly, it is left out in the Geno part, leading to a vector of same dimensionality as the Geno representation itself.

### Combining individual-drug scores into regimen scores

Traditionally, the activity of a drug combination has been determined from the activities of the individual components by simple summation. This approach has been termed genotypic/phenotypic susceptibility (or: sensitivity) score (GSS/PSS) (DeGruttola et al., 2000) and has been used for the approaches (1–4). A more recent study has refined the summation by using drug-specific weights when combining the individual drug activities. Here, we compare two different ways of combining individual-drug scores into a regimen score: summation (for inputs 1–4); statistical learning (for all inputs) using support vector classification. Support vector classification was performed using a Gaussian radial basis function (RBF) kernel. The RBF parameter  $\gamma$  was used with *libsvm* default value, class weights were set to adjust for the class skew between successes and failures, and the soft margin penalty  $C$  was optimized for each input as described below. A continuous score normalized within the interval  $[0, 1]$  was obtained by Platt’s probabilistic method for SVMs (Platt, 1999).

### Evaluation setup

Models were compared by 10-fold cross-validation. To ensure an unbiased comparison, the model selection procedure was integrated into the evaluation procedure. Variants of this approach are sometimes referred to as “double” or “nested” cross-validation (Ruschhaupt et al., 2004). The goal is to avoid overfitting hyperparameters which could happen in simple cross-validation. Specifically, the three evaluations reported here were performed as follows:

1. **Genotype-centric standard datum; cross-validation on Stanford-Kaiser data.** 1,000 random TCEs were drawn from the Stanford-Kaiser data (consisting of 6,337 TCEs) to determine the optimal value of  $C$ . The set was randomly

split into 750 training and 250 validation samples. For each of the nine input alternatives introduced above, and for each  $C \in \{2^{-1}, 2^{-0.5}, 2^0, 2^{0.5}, \dots, 2^6\}$ , a support vector classifier was fitted to the training data. The models were then used to predict the 250 validation samples and for each of the nine input alternatives, the  $C$  with the highest area under the ROC curve was chosen as the best hyperparameter. The remaining 5,337 TCEs were then used with the optimal  $C$  value to obtain an unbiased estimate of model performance using ten-fold cross-validation. This whole procedure was repeated ten times to account for randomness in determining the optimal value of  $C$ . Thus, for each of the nine inputs, ten (possibly different values of  $C$ ) times ten (cross-validation) sets of “blind” predictions, each of size 533 or 534, were available for performance evaluation. For each of these sets, we also determined regimen scores by the simple summation method for comparison.

- 2. Genotype-centric standard datum; training on Stanford-Kaiser data, prediction on EuResist data.** Even though all measures have been taken to prevent overfitting, it is obvious that models heavily depend on the properties of the data from which they are fitted. It has been observed previously that the distribution of used drug combinations is heavily skewed ([Beerenwinkel, 2004](#)), reflecting approval times, changes in treatment strategies, or even regional differences from hospital to hospital ([Larder et al., 2007](#)). To estimate how well our results generalize not only across data similar to that collected in the Stanford-Kaiser effort, but across completely different collection efforts, we performed this second analysis. For each of the nine inputs, and for each of the ten values of  $C$  determined as optimal for a particular input in the first analysis, models were trained on the full Stanford-Kaiser data and then used to predict the full EuResist data. Thus, for each of the nine inputs, ten sets (possibly different values of  $C$ ) of “blind” predictions, each of size 5,224, were available for performance evaluation.
- 3. Classical standard datum; cross-validation on Stanford-Kaiser data.** The genotype-centric standard eliminates the evolutionary forces behind therapy failure and focuses on the role of the genotype at time of failure. Thus, it is well-suited to study how big the alleged dichotomy between clinical and genotypic drug resistance really is. Moreover, it can be, and in fact is being used in practice to rank combination therapies by their probability of success ([Altmann et al., 2007](#)). Still, by its retrospective look on failures, it does not mimic the real situation when a physician has to choose a novel therapy without knowledge about a future genotype. Therefore, we repeated the first analysis using the classical standard datum. The same ten values of  $C$  as determined in analysis (1) were used, so that 10-times cross-validation could be performed on the whole data. Thus, for each of the nine inputs, ten times ten sets of “blind” predictions, each

of size 235 or 236, were available for performance evaluation. For each of these sets, we also determined regimen scores by the simple summation method for comparison.

### Measures of predictive performance

The output of both the simple summation, as well as the support vector machine, is given by a score rather than an actual class prediction. The analysis of such scoring classifiers is usually performed in the framework of ROC analysis (Fawcett, 2006). The analyses were performed using the classifier evaluation package ROCR (Sing et al., 2005). We considered successes as “positive” and failures as “negative” samples, and focused on the following five performance measures:

- $TPR_{10}$  and  $TPR_{20}$ : The true positive rate (sensitivity, recall) of a scoring classifier at a cutoff the induces a false positive rate of 10% (equivalently, a specificity of 90%) or of 20%.
- AUC,  $AUC_{10}$  and  $AUC_{20}$ : The area under the ROC curve (providing a cutoff-independent measure of class separation by a scoring classifier), and the partial AUCs up to a false positive rate of 10% and 20%, respectively.

### 5.3.3 Results

#### Regimen scores from individual drug scores: summation versus statistical learning

Our first analysis was devoted to comparing summation versus statistical learning as methods for combining scores for single drugs into regimen scores. Figure 5.8 shows averaged ROC curves (10 replicates of 10-fold cross-validation) based on the ANRS, Rega, HIVDB, and Pheno input representation, respectively. For each representation, regimen scores are either derived by simple summation or by statistical learning, as described in Section 5.3.2. The graph clearly shows that at all practically relevant false positive rates, the regimen score based on statistical learning drastically outperforms the summation-based regimen score. Table 5.2 shows the performance improvements conferred by statistical learning according to a number of performance criteria. For all input representations and performance criteria, the improvements are consistent and highly statistically significant ( $p < 2.2 \times 10^{-16}$ , Wilcoxon signed rank test). For example, at a false positive rate of 10%, the improvements in true positive rate range from 19.2% (ANRS) to 31.3% (Pheno). Given these clear results, all further experiments were performed with regimen scores derived by statistical learning.

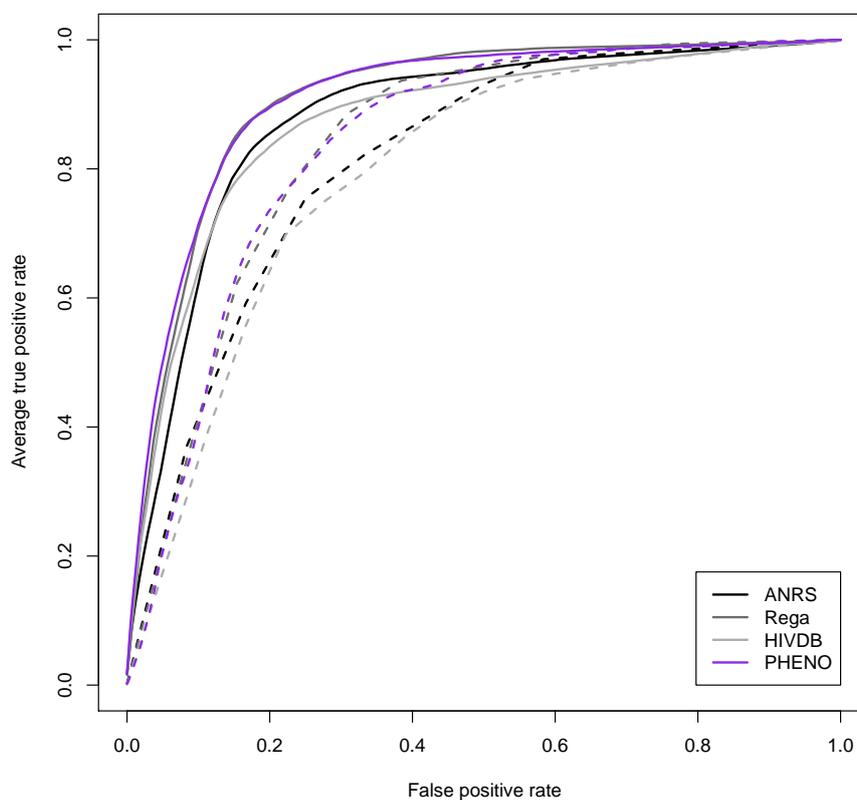


Figure 5.8: Comparing summation (dashed lines) vs. statistical learning (solid lines) for combining individual-drug scores into regimen scores.

Table 5.2: Mean (standard error) performance improvement (across 10 replicates of 10-fold cross-validation) for selected performance measures when using statistical learning rather than simple summation to combine individual drug scores into regimen scores.  $AUC_{10}$  refers to the AUC up to a false positive rate of 10%.

	$AUC_{10}$	$AUC_{20}$	AUC	$TPR_{10}$	$TPR_{20}$
ANRS	0.013 (0.005)	0.034 (0.009)	0.063 (0.016)	0.192 (0.08)	0.19 (0.056)
Rega	0.022 (0.005)	0.045 (0.009)	0.065 (0.015)	0.291 (0.069)	0.177 (0.044)
HIVDB	0.022 (0.006)	0.048 (0.009)	0.078 (0.017)	0.283 (0.066)	0.176 (0.048)
PHENO	0.026 (0.006)	0.048 (0.009)	0.071 (0.016)	0.313 (0.076)	0.159 (0.049)

## Method comparison

The second experiment was devoted to a detailed comparison of the five non-hybrid input representations and the four hybrid input representations combining a genotypic representation (ANRS, Rega, HIVDB, Geno) with the Pheno representation. Averaged ROC curves from 10 replicates of 10-fold cross-validation on the Stanford data are shown in Figure 5.9(a), and numerical readouts of the graph at selected false positive rates are shown in the upper part of Table 5.3. Remarkably, the Pheno representation is among the best non-hybrid representations overall, with the highest  $AUC_{20}$  and AUC. In terms of  $AUC_{10}$  and  $TPR_{10}$ , the Geno representation is the best non-hybrid representation, and in terms of  $TPR_{20}$  the Rega representation is best. As can be seen in Figure 5.9(a), even the worst hybrid representation outperforms the best non-hybrid representation at all practically relevant false positive rates. Table 5.4 shows p-values of comparisons (Wilcoxon signed rank test) between the best non-hybrid and the worst and best hybrid representation, respectively. For all performance measures but  $AUC_{10}$ , the improvement conferred even by the worst hybrid representation is significant (first row). When comparing the best non-hybrid with the best hybrid representation (second row), the improvements are highly significant for all performance measures.

Data sets collected from different regions at different times may be very different from each other, reflecting local or temporal changes in preferred treatment strategies. In order to investigate to which degree classifiers learned from the Stanford data can be used for prediction on completely different data, we used them for predicting the EuResist data set, which was not used at all during training. ROC curves are shown in Figure 5.9(b), and selected numerical readouts are presented in the lower part of Table 5.3. Generally, the prediction performance of all methods decreases when used on the completely different data. For example, on the Stanford data, the best methods achieve a  $TPR_{10}$  of 76.3% and a  $TPR_{20}$  of 92.7% (in cross-validation), compared to 53.4% and 74.3% when trained on the Stanford, but evaluated on the EuResist data. However, qualitatively, the results remain similar in that again hybrid methods outperform non-hybrid methods across all relevant false positive rates.

## Classical standard datum

As a final experiment, we replicated the two experiments reported above using the classical standard datum definition instead of the genotype-centric standard datum. The ROC curves (from ten replicates of ten-fold cross-validation) in Figure 5.10(a) are the results of comparing summation versus statistical learning for combining single-drug scores into regimen scores. Similar to the genotype-centric standard datum, statistical learning outperforms summation for all the methods investigated, although the differences are smaller than in the alternative case.

Figure 5.10(b) shows the results – now using the classical standard datum – of

Table 5.3: Selected measures of model performance for hybrid and non-hybrid input representations.

Cross-validation on Stanford-Kaiser data.					
	AUC <sub>10</sub>	AUC <sub>20</sub>	AUC	TPR <sub>10</sub>	TPR <sub>20</sub>
ANRS	0.035 (0.005)	0.112 (0.009)	0.879 (0.017)	0.622 (0.077)	0.855 (0.037)
Rega	0.042 (0.005)	0.125 (0.009)	0.907 (0.013)	0.708 (0.06)	0.898 (0.031)
HIVDB	0.04 (0.005)	0.116 (0.008)	0.873 (0.017)	0.644 (0.064)	0.835 (0.031)
Pheno	0.046 (0.006)	0.129 (0.009)	0.908 (0.015)	0.714 (0.061)	0.895 (0.035)
Geno	0.047 (0.006)	0.129 (0.008)	0.908 (0.014)	0.725 (0.052)	0.886 (0.031)
ANRS+Pheno	0.048 (0.006)	0.134 (0.008)	0.92 (0.012)	0.761 (0.054)	0.92 (0.025)
Rega+Pheno	0.048 (0.006)	0.133 (0.009)	0.92 (0.013)	0.744 (0.053)	0.923 (0.03)
HIVDB+Pheno	0.05 (0.006)	0.135 (0.008)	0.922 (0.012)	0.761 (0.049)	0.927 (0.024)
GENO+Pheno	0.048 (0.006)	0.134 (0.008)	0.915 (0.013)	0.763 (0.051)	0.906 (0.029)
Training on Stanford-Kaiser data, prediction on EuResist data.					
	AUC <sub>10</sub>	AUC <sub>20</sub>	AUC	TPR <sub>10</sub>	TPR <sub>20</sub>
ANRS	0.019 (0.002)	0.066 (0.005)	0.802 (0.006)	0.318 (0.03)	0.608 (0.007)
Rega	0.028 (0.003)	0.089 (0.004)	0.843 (0.007)	0.499 (0.036)	0.703 (0.01)
HIVDB	0.025 (0.004)	0.076 (0.004)	0.781 (0.009)	0.4 (0.032)	0.622 (0.013)
Pheno	0.028 (0.002)	0.09 (0.005)	0.841 (0.007)	0.481 (0.029)	0.732 (0.014)
Geno	0.037 (0.001)	0.099 (0.002)	0.84 (0.002)	0.525 (0.006)	0.716 (0.005)
ANRS+Pheno	0.031 (0.001)	0.095 (0.002)	0.859 (0.003)	0.506 (0.016)	0.739 (0.006)
Rega+Pheno	0.031 (0.003)	0.094 (0.005)	0.857 (0.007)	0.519 (0.03)	0.737 (0.011)
HIVDB+Pheno	0.034 (0.005)	0.099 (0.007)	0.862 (0.009)	0.534 (0.043)	0.743 (0.012)
Geno+Pheno	0.033 (0.003)	0.097 (0.004)	0.846 (0.004)	0.528 (0.023)	0.74 (0.009)

Table 5.4: Statistical significance (logarithm of the p-values) of performance differences observed in cross-validation on the Stanford data.

	AUC <sub>10</sub>	AUC <sub>20</sub>	AUC	TPR <sub>10</sub>	TPR <sub>20</sub>
Best non-hybrid	Geno	Pheno	Pheno	Geno	Rega
Worst hybrid	-0.87	-9.92	-8.28	-3.42	-2.21
	ANRS+Pheno	Rega+Pheno	Geno+Pheno	Rega+Pheno	Geno+Pheno
Best hybrid	-4.20	< -15	< -15	-9.54	< -15
	HIVDB+Pheno	HIVDB+Pheno	HIVDB+Pheno	Geno+Pheno	HIVDB+PHENO

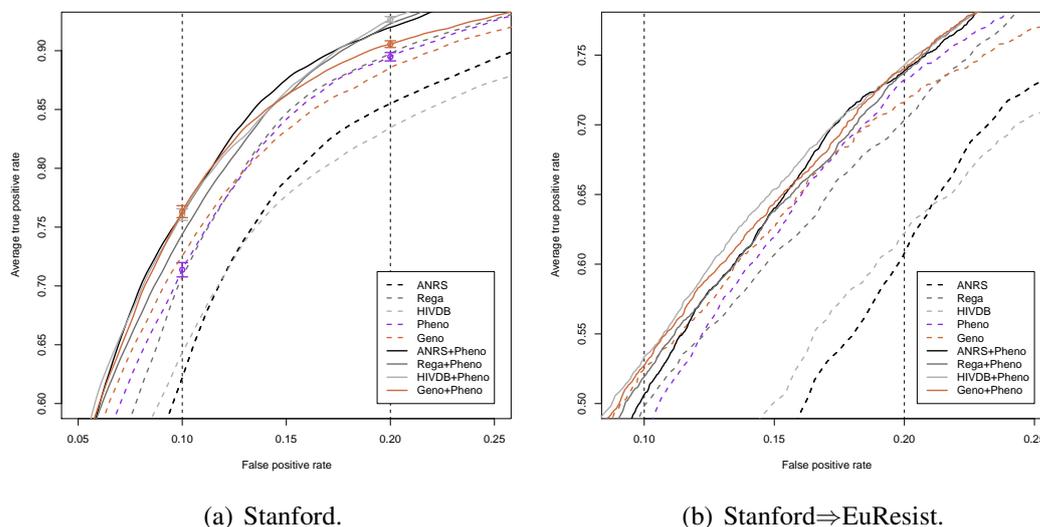


Figure 5.9: Comparing hybrid (solid lines) and non-hybrid (dashed lines) input representations.

comparing the five non-hybrid input representations and the four hybrid input representations that combine a genotypic representation (ANRS, Rega, HIVDB, Geno) with the Pheno representation. With the general decrease in predictive performance, the advantage of hybrid methods is greatly reduced, as compared to the genotype-centric standard datum. Remarkably, as shown in Table 5.5, the Pheno representation achieves the highest AUC among all non-hybrid representations.

### 5.3.4 Discussion

In this large-scale study, we have compared a variety of input representations for their ability to discriminate between treatment success or failure, based on genotype and a chosen drug combination. Two different methods for combining scores for individual drugs into scores for a regimen as a whole have been evaluated, based on either the traditional summation, or on statistical learning using support vector machines. The use of statistical learning for score combination consistently outperformed the summation in our experiments. We hypothesize that the superiority of the learning-based combination is due to its intrinsic ability to assign specific weights for individual drugs in a regimen (possibly corresponding to potency), in contrast to the traditional unweighted summation, and for capturing drug-drug interaction effects, as far as they can be learned from the data. The benefits of weighting the contributions of individual drugs has previously been reported in the context of a linear model (Swanstrom et al., 2004).

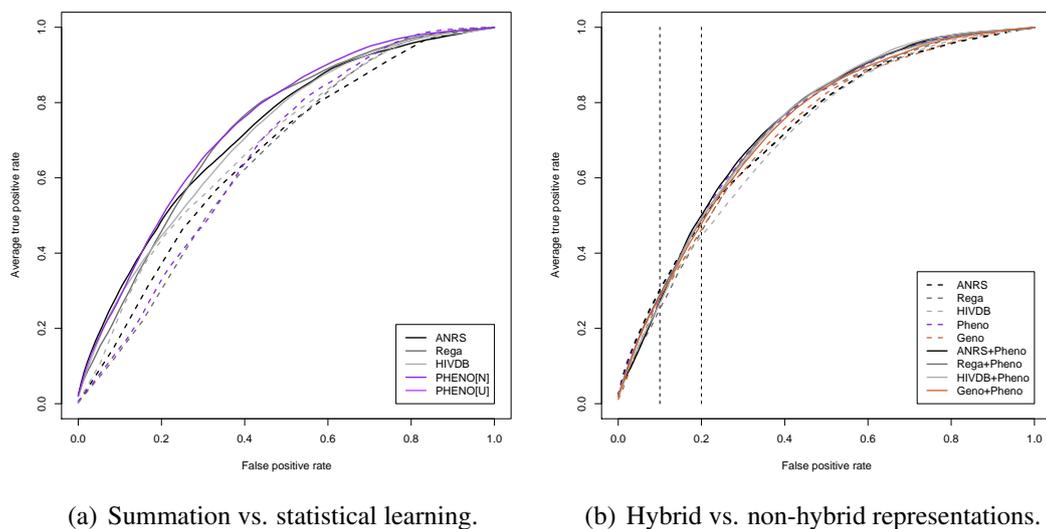


Figure 5.10: Replication of the previous experiments, now based on the classical, rather than the genotype-centric standard datum.

Table 5.5: Selected measures of model performance, estimated by cross-validation on the Stanford data, using the classical standard datum.

	AUC	TPR <sub>10</sub>	TPR <sub>20</sub>
ANRS	0.723 (0.033)	0.303 (0.075)	0.486 (0.074)
Rega	0.727 (0.036)	0.252 (0.075)	0.458 (0.094)
HIVDB	0.712 (0.034)	0.286 (0.073)	0.446 (0.065)
Pheno	0.74 (0.033)	0.283 (0.068)	0.495 (0.084)
Geno	0.719 (0.033)	0.273 (0.067)	0.463 (0.084)
ANRS+Pheno	0.741 (0.033)	0.271 (0.079)	0.5 (0.082)
Rega+Pheno	0.74 (0.034)	0.29 (0.067)	0.488 (0.085)
HIVDB+Pheno	0.739 (0.035)	0.267 (0.078)	0.492 (0.087)
Geno+Pheno	0.733 (0.031)	0.283 (0.066)	0.478 (0.078)

A major focus of our analysis was to compare the performance of input representations directly based on the genotype with an input representation based on predicted phenotypes. As mentioned in Section 5.3.1, this latter approach has been criticized frequently, although there has been little data as a basis for judgement. For example, it has been mentioned that "attempting to infer response from genotype via the intermediate step of predicting the phenotype is likely to have serious limitations" (Larder et al., 2007; Brun-Vézinet et al., 2004). Contrary to this widely held belief, none of the experiments reported here shows any limitation of the phenotype-based representation, as compared to the genotype-based representations. Rather, according to several performance measures (AUC, AUC<sub>20</sub>), the phenotype representation is even the best non-hybrid input representation. It is also in closest correspondence with an actual measurable quantity among all input representations, namely the resistance factor (fold-change in IC<sub>50</sub>). It is important to keep in mind that the phenotype-based representation relies on normalization for realizing its full potential (e.g. in the way described in Section 5.3.2). This is most likely due to the very different range of resistance factors observed for the different drugs, and their nonuniform clinical relevance.

A third focus of our study was to assess potential synergies between genotypic representations and the phenotypic representation. Intuitively, a certain degree of complementarity might be expected between representations that were derived from very different kinds of data. Moreover, approaches based on integrating these various sources of data could lead to more robust discriminations between successes and failures. Confirming this hypothesis, a substantial and highly significant benefit was observed for hybrid over non-hybrid methods in cross-validation experiments on the Stanford data. This benefit was qualitatively preserved when using classifiers trained on the Stanford data for predicting samples from the EuResist data. However, the difference was much smaller when using the methods on a completely different data set. Finally, when using the classical, instead of the genotype-centric standard datum, the benefits of hybrid over non-hybrid representations have virtually vanished. We were not able to explain these dataset-dependent discrepancies. Still, when using the genotype-centric standard datum, such as for ranking therapies (as currently done in the THEO system<sup>12</sup>), the use of a hybrid representation is likely to contribute to an increased predictive performance.

In summary, we have shown that the use of a phenotypic input representation is clearly competitive with genotype-based representations, the former often outperforming the latter. Moreover, substantial synergies can exist between these different representations, which can be explained by the different kinds of data from which they are derived. However, in our experiments, the magnitude of these synergies strongly depends on the data set and on the standard datum definition. For a better understanding and definition of the potential benefits of hybrid approaches, these and similar exper-

---

<sup>12</sup>Available at [www.geno2pheno.org](http://www.geno2pheno.org).

iments should be replicated with different datasets, standard datum definitions, and possibly more sophisticated forms of hybrid representations.

## 5.4 Inferring response from genotype: with or without predicted replication capacity?<sup>13</sup>

In Section 5.3, we have shown how the incorporation of predicted resistance phenotypes can improve the reliability of response prediction. In this short section, we study the clinical relevance of another phenotypic quantity of potential clinical relevance: viral replication capacity (RC), or “fitness”<sup>14</sup>. The derivation of genotype-RC models using support vector machines and random forests is described in detail in (Weisser, 2006; Sing et al., 2006). Here, we only briefly review the necessary background from these studies, but focus on the previously unreported evaluation of the clinical relevance of predicted RC.

### 5.4.1 Introduction

Although there is a correlation between drug resistance and virologic response, development of resistance does not always entail treatment failure. Many of these discordant cases might be explained by non-viral effects, such as host or pharmacological factors. However, in several cases, the discordance was linked to the presence of specific mutations in the viral genome. A well-known example is the M184V mutation which confers high-level resistance to lamivudine and emtricitabine. Still, despite resistance, viral load often remains below baseline levels in patients harboring M184V strains.

Based on the reported cases of discordance, the clinical use of phenotypic resistance data has been said to suffer from “serious limitations as the association between *in vitro* resistance levels and virological response is often not well characterized” (Brun-Vézinet et al., 2004). As an alternative, it has been suggested “to refine and optimize the use of genotyping by establishing interpretative tools or algorithms based on direct correlations [of genotypes] with virological response” (Brun-Vézinet et al., 2004).

However, compared to phenotypic measurements which are obtained from well-controlled and reproducible assays, virological response is influenced by many additional factors. Moreover, models based on the direct correlation between genotype and response do not even try to understand the causal relationships involved in determining

---

<sup>13</sup>Part of the work reported in this section (on genotype-fitness data) was performed in collaboration with Hendrik Weisser (MPI for Informatics, Saarbrücken), in the context of a Bachelor’s thesis, and with Hauke Walter and Monika Tschochner (University of Erlangen).

<sup>14</sup>Due to the many different uses of the word “fitness”, we prefer to use the technical term “replication capacity”.

response to therapy, but rather treat virus-host interaction entirely as a “black box”.

For these reasons, it is worthwhile further exploring the use of phenotypic information (typically predicted from genotype) in predicting response to therapy. Thus, while the “black box” paradigm mentioned above concludes that “less modeling” (in the sense of discarding phenotypic information and directly correlating genotype to response) is the appropriate reaction to reported resistance-response discordance, we suggest it should be “more modeling”. For example, we have previously shown that the integration of evolutionary information into models of therapy response improves the reliability of the predictions (Altmann et al., 2007; Sing and Beerenwinkel, 2007).

In the present study, we explore the use of another factor of potential relevance: viral replication capacity (RC). In several cases in which specific mutations have been implicated with discordant phenotypic and *in vivo* behavior, this was attributed to RC effects. Thus, genotype-based approaches for predicting response to therapy could potentially benefit from integrating predicted RC.

## 5.4.2 Materials and Methods

**Genotype-RC data.** Two datasets of matched genotype-RC pairs were available for building genotype-RC models. Dataset “Erlangen” consisted of 261 samples measured using a novel RC assay developed at the University of Erlangen-Nürnberg (reviewed in Weisser, 2006). Genotyping was performed for the whole protease, and for residues 1–250 of the reverse transcriptase. Dataset “Virologic/Monogram” consisted of 317 samples measured using the commercial Monogram replication capacity assay<sup>15</sup>. Here, genotypes were available for protease positions 4 to 99 and for reverse transcriptase positions 38 to 223. Genotype-RC models were derived by support vector regression and random forest regression, with model selection and performance evaluation as described in detail in (Weisser, 2006). In particular, RC was represented as the decadic logarithm of the measured RC. For example, a measured RC value of 70% would be represented as  $\log_{10}(70) \approx 1.85$ .

**Clinical genotype-therapy-response data.** For analyzing clinical relevance, a subset of the “Stanford-Kaiser” data described in Section 5.3 were available. These data were dichotomized into success and failure according to the “alternative” standard datum (cf. Section 5.3). Moreover, the data had been “balanced on therapies”, meaning that for each drug combination the number of successful treatment change episodes (TCEs) was equal to the number of failing TCEs. In this way, no drug will appear superior *per se* compared to any other drug in the statistical learning procedure. This is in contrast to the unbalanced datasets used in Section 5.3, where for example the ratio of successful to failing therapies is much higher for lopinavir-containing regimens than

---

<sup>15</sup>Kindly provided by Mark Segal (University of California, San Francisco).

for zidovudine-containing regimens. In total, there were 1,022 successful TCEs and the same number of failing TCEs.

**Evaluation of clinical relevance.** To assess the clinical relevance of predicted RC, we evaluated three different input representations. In representation RC, each of the 2,048 samples was simply represented by the RC (predicted with an SVM model derived from the “Virologic/Monogram” data) and the drugs in the regimen, leading to a (1 [RC] + 17 [drugs])-dimensional vector for each sample. Representation RF (for “resistance factor”) was exactly the “Pheno” representation from Section 5.3, in which a sample was represented by a 17-dimensional vector representing drug resistance for the samples in the regimen. Finally, representation RF+RC consisted of representation RF, to which the predicted RC was concatenated, leading again to a (17 [RF/drugs] + 1 [RC])-dimensional representation for each sample. Each of these three input representations was then correlated to response (success or failure) using support vector classification with a linear kernel. Ten replicates of ten-fold cross-validation were used to assess predictive performance.

### 5.4.3 Results

**Predictive reliability of the genotype-RC models.** In 10-fold cross-validation, the squared correlation coefficient  $r^2$  of support vector regression attained a mean of 0.241 (Erlangen,  $\sigma = 0.015$ ) and 0.316 (Monogram,  $\sigma = 0.018$ ). Random forests led to a mean  $r^2$  of 0.297 (Erlangen,  $\sigma = 0.019$ ) and 0.272 (Monogram,  $\sigma = 0.028$ ). This predictive performance is substantially lower than that observed for most drugs when predicting phenotypic resistance, even in datasets of the same size, and using exactly the same methods. We could not identify the reasons why reliable prediction of RC was not possible here. For the Erlangen data set, repeated measurements of one sample were available to investigate the assay variability (these repeated measurements were not used in training the classifiers). As shown in Figure 5.11, the assay variability is substantial, spanning almost 50% of the range of fitness values between 0 and 100%. There were no repeated measurements available to us for the Monogram dataset, so no comparison was possible. Certainly, a measurement noise of this degree is likely to exhibit a strong burden on the expected predictive performance.

**Evaluation of clinical relevance.** Figure 5.12 shows the results of evaluating the different input representations in the task of predicting success or failure of antiviral combination therapy. The yellow curve, representing the predictive performance of the RC representation, is close to the diagonal for the practically relevant false positive rates. The combined representation RF+RC outperforms the RF representation only marginally.

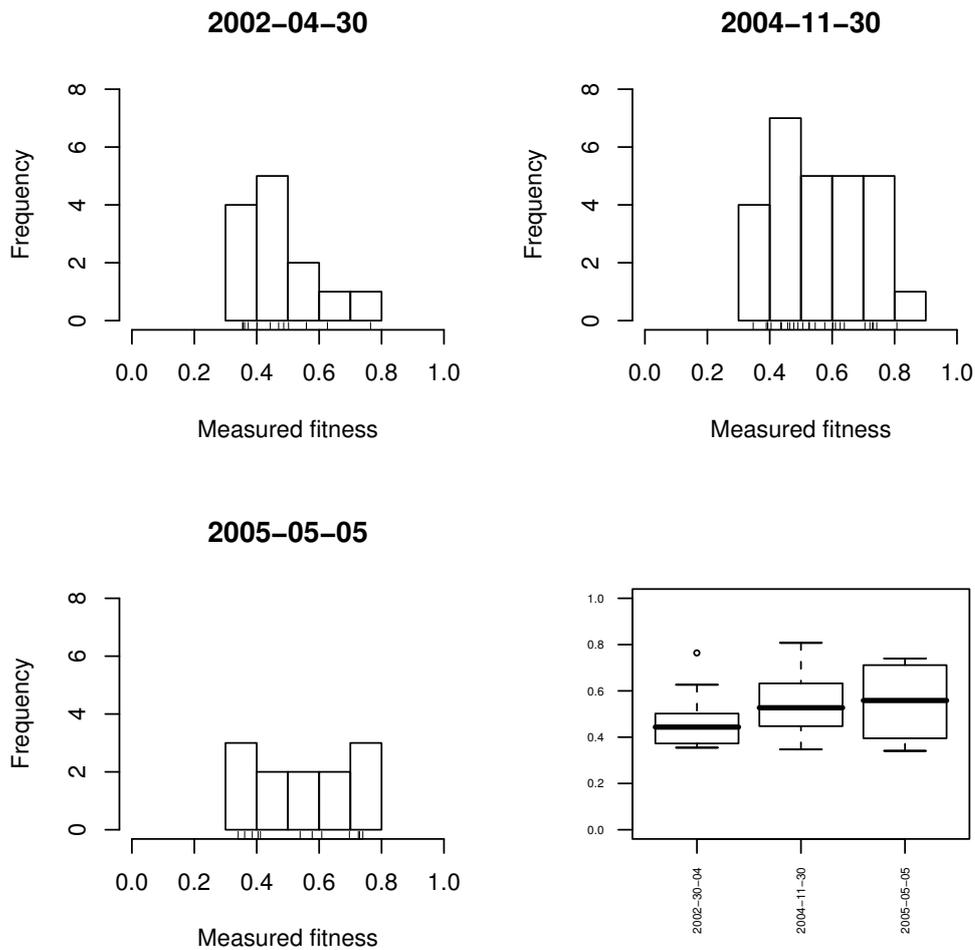


Figure 5.11: Repeated measurements of a single sample at three different dates, showing the substantial assay variability and a slight trend over time.

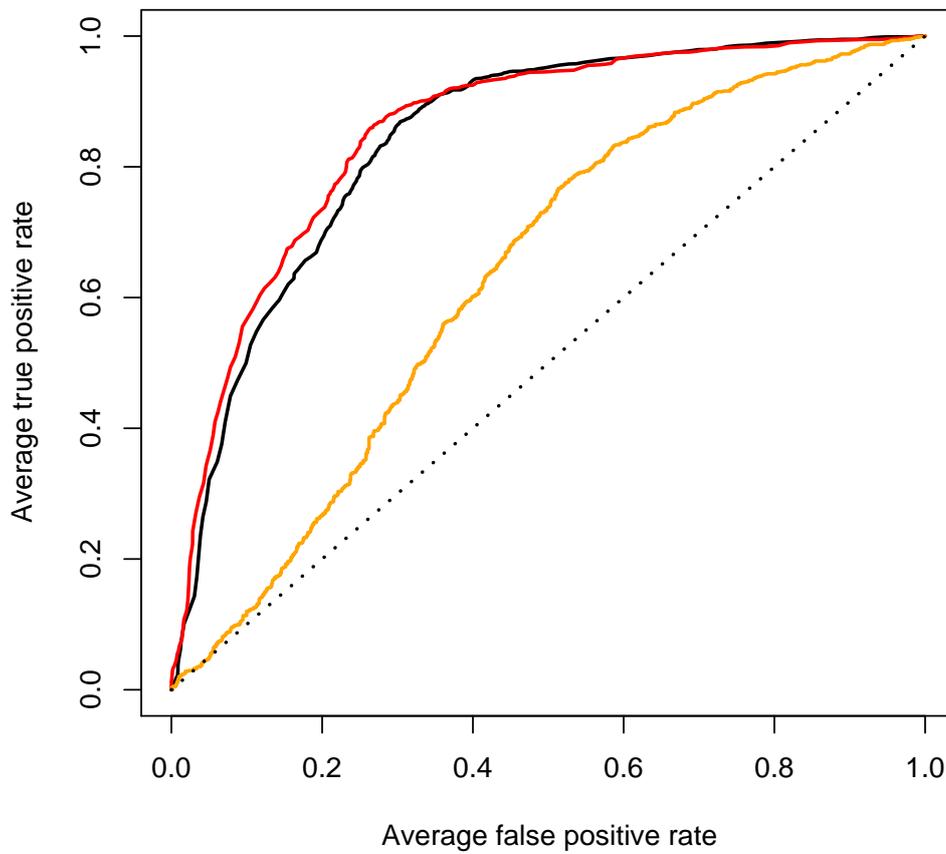


Figure 5.12: Averaged ROC curves (from 10 replicates of ten-fold cross-validation) for the prediction of success or failure of combination therapy. The dotted diagonal represents completely uninformed prediction. The yellow, black, and red curves represent classification purely based on RC, purely based on RF, and based on both RF and RC, respectively.

#### 5.4.4 Discussion

In this section, we have studied the integration of predicted replication capacity into a classifier for predicting success or failure of antiviral combination therapy. The observed benefits were marginal, compared to an approach based exclusively on predicted phenotypic drug resistance. Thus, in the current form, information about viral replication capacity cannot be used to improve systems such as the therapy ranker THEO<sup>16</sup>. We hypothesize three alternative explanations for the lack of utility of RC. First, the predictive performance of genotype-RC regression was substantially below that observed in the prediction of phenotypic resistance. The predicted RCs might be too noisy to be useful for other purposes. Second, it might be that the naive integration of RC with RF by simple concatenation was not adequate, and that more sophisticated forms of integration might be needed. Third, as of today, the clinical relevance of replication capacity is not fully established (De Luca, 2006), and the lack of clinical utility of predicted RC might be a symptom of a lack of clinical utility of RC in general. Additional studies will be needed in order to identify which of these three alternatives is the true explanation for our observations.

---

<sup>16</sup>Available at [www.geno2pheno.org](http://www.geno2pheno.org).

---

# Genotypic, Clinical, and Structural Predictors of HIV Coreceptor Usage

HIV coreceptor usage is tightly linked to disease progression towards AIDS. Moreover, drugs from the novel class of coreceptor antagonists aim to selectively inhibit one of the two main coreceptors for HIV, called CCR5 and CXCR4. For these two reasons, it is important to understand the relation between mutations in the viral genome and changes in the coreceptor usage of the viral population. This chapter is structured into four main parts. First, we shall review the clinical motivation for predicting coreceptor usage, including a brief account of the exciting history of research on coreceptor usage (Section 6.1). Section 6.2 is devoted to analyzing the relation between genotype and coreceptor usage phenotype on both clonal and clinical (“population-based”) data, including a discussion of related work. Because (as we shall see) clinical data is “harder” to deal with than clonal data, we devote Section 6.3 to the incorporation of sequence, clinical and host data only available in a clinical setting. Finally, in Section 6.4, we show how the incorporation of structural modeling into a prediction system can improve both predictive performance as well as our understanding of the genotype-phenotype relation.

## 6.1 Introduction

### Motivation

HIV-1 enters target cells through a multi-stage interaction of the viral envelope protein gp120 with the CD4 host cell receptor and a cellular coreceptor, usually CCR5 or CXCR4 (reviewed in [Sirois et al., 2005](#)). Individual virions are able to use one or the other or both coreceptors (R5/X4 phenotype). *In vivo*, R5-only virus is generally present over the entire course of infection ([Schuitemaker et al., 1991](#)), and CXCR4-capable (R5/X4 or X4) variants are detected in approximately 50% of infected individuals at end-stage disease. The reason for this coreceptor switch remains unclear,

but both *in vitro* studies and experiments in animal models suggest that the emergence of X4 virus is strongly associated with CD4<sup>+</sup> cell depletion and thus may be an important determinant of pathogenesis (Regoes and Bonhoeffer, 2005). The question of whether X4 virus is a cause or emerges as a result of CD4<sup>+</sup> cell depletion (or both) as well as the evolutionary reasons for the development of these variants, remain largely unresolved (Penn et al., 1999). The capacity of the virus to use CXCR4 lies at least partially in a change of several amino acids in the third hypervariable (V3) loop of gp120 (Fouchier et al., 1992), although sequence changes outside the V3 region also contribute to coreceptor use (Groenink et al., 1992).

In recent years, substantial attention has been devoted to HIV coreceptors due to their potential as drug targets, and antagonists of the CCR5 coreceptor are now in advanced clinical studies (Dorr et al., 2005; Westby and van der Ryst, 2005). Notably, coreceptor antagonists represent the first class of anti-HIV drugs targeting a host protein, rather than a viral protein. Unfortunately, the static nature of this target (as opposed to the rapidly adapting viral enzymes targeted by classical drugs), does not prevent the emergence of resistant mutants. Specifically, resistance to coreceptor antagonists can include increased viral binding affinity to the coreceptor, changes in the CCR5 binding mode, or the emergence of newly produced or pre-existing X4 variants (Moore et al., 2004). Given the link between emerging X4 virus and disease progression, the need for careful monitoring of viral coreceptor usage for screening and treatment with CCR5 antagonists becomes apparent.

Phenotypic assays for monitoring viral drug resistance or coreceptor usage are commercially available, although they are relatively expensive and have a relatively slow turnaround. Approaches based on the viral genotype promise potential alternatives for routine clinical usage. The complex relationship between viral genotype and phenotype and/or response to therapy has led to the development of sophisticated interpretation algorithms which have been successfully implemented for HIV drug resistance testing. As reviewed in Section 5.1, these algorithms are now widely used to support treatment with the antiretroviral drug classes of protease (PIs) and reverse transcriptase inhibitors (RTIs). Genotypic approaches for monitoring coreceptor usage aim at detecting X4-capable virus with high sensitivity, while minimizing the number of false positives, i.e. R5-only variants that are incorrectly predicted as X4-capable. To date, the most popular genotypic predictor of X4-capable virus is the simple 11/25 rule, predicting X4-capable virus based on the presence of arginine or lysine at positions 11 and/or 25 of the third hypervariable (V3) region of the envelope protein gp120 (Fouchier et al., 1992; de Jong et al., 1992; Fouchier et al., 1995; Korber et al., 1993). However, previous studies have found the overall reliability of sequence motif-based methods for phenotype inference, especially for coreceptor usage prediction, to be limited (Resch et al., 2001). Moreover, it has been suggested that for many V3 backgrounds, basic changes at position 11 or 25 are neither necessary nor sufficient for a phenotype switch (Jensen et al., 2003). As a consequence, several alternatives based on

statistical learning methods have been developed (reviewed in [Jensen and Wout, 2003](#)), including linear regression ([Briggs et al., 2000](#)), artificial neural networks ([Resch et al., 2001](#)), decision trees ([Pillai et al., 2003](#)), support vector machines ([Pillai et al., 2003](#)), position-specific scoring matrices ([Jensen et al., 2003](#)), and mixtures of localized rules ([Sing et al., 2004](#)).

### A brief history of research on HIV coreceptor usage

In this section, we briefly review the events that led to our current understanding of HIV coreceptor usage and the implication of the gp120 V3 loop as a major determinant. Among other factors, this history is remarkable because of its delayed onset: Reverse transcriptase inhibitors had been in use already for a decade and the first protease inhibitors had just been approved when the first reports appeared showing that HIV was dependent on a coreceptor to enter a host cell.

The fact that HIV-1 could only infect cells with the CD4 receptor was discovered very early ([Dalglish et al., 1984](#); [Klatzmann et al., 1984](#)), in fact almost immediately after the discovery of HIV-1 ([Barré-Sinoussi et al., 1983](#); [Popovic et al., 1984](#)). However, research on HIV coreceptors began over a decade after these discoveries. First, [Cocchi et al. \(1995\)](#) observed that three chemokines, MIP1- $\alpha$ , MIP1- $\beta$ , and RANTES (nowadays systematically called CCL3 to CCL5) were potent repressors of strains with *in vitro* tropism for macrophages. Next, [Feng et al. \(1996\)](#), discovered that some strains infecting T-cells *in vitro* needed a particular chemokine receptor – which we call CXCR4 today – as a coreceptor, in addition to CD4. However, the three chemokines did not bind to CXCR4. Due to this observation it was widely hypothesized that CCL3 to CCL5 were the natural ligands of a hypothesized “second” coreceptor which is needed by some viral variants as a coreceptor. Shortly afterwards, this receptor, today called CCR5, was identified simultaneously by several groups ([Alkhatib et al., 1996](#); [Deng et al., 1996](#); [Dragic et al., 1996](#); [Choe et al., 1996](#); [Doranz et al., 1996](#)).

At the very moment of their discovery, the idea of inhibiting the HIV coreceptors to prevent viral entry was born. Furthermore, the discovery of differential coreceptor usage also replaced previous phenotype classification systems based on cell tropism, replication rate in peripheral blood mononuclear cells (PBMCs), or the cytopathology in MT-2 cells ([Fenyo et al., 1997](#)). Based on these “old” classification schemes, which are highly correlated but not identical to coreceptor usage, the third variable region of gp120 had already been identified as a major determinant of phenotype ([Chesebro et al., 1991](#); [Hwang et al., 1991](#)). The rough region being identified, interest grew in determining the relation between genotype and phenotype more precisely. Two pioneering papers appeared in 1992, implicating a net V3 charge of at least five ([Fouchier et al., 1992](#)) or the presence of a basic residue at V3 positions 11 or 25 ([de Jong et al., 1992](#)) with usage of the CXCR4 receptor. At this time it was already known that a phenotype switch was associated with progression to AIDS ([Tersmette et al., 1989](#)).

Thus, a direct link between the sequence evolution of V3 and disease progression had been established. Reflecting the growing interest in this link, the first article on V3 evolution appeared soon afterwards (Kuiken et al., 1992). Long before the V3 region was implicated as a phenotype determinant, it had been recognized as a major target for neutralizing antibodies. In an early bioinformatics paper by Modrow et al. (1987), a region within V3 was predicted as an epitope, along with other regions. This prediction was confirmed in the following year (Goudsmit et al., 1988), and in 1989, the V3 region was given the name the "principal neutralizing determinant", because its deletion stopped the activity of neutralizing antibodies. Thus, in 1996, all questions were there which have occupied HIV coreceptor research until today, and will continue to do so for a while.

## 6.2 Predicting HIV coreceptor usage from sequence<sup>1</sup>

### 6.2.1 Related work

Ultimately, all information for determining the coreceptor usage of a given virus strain is contained in its genome. Thus, the relation between sequence alterations and changes in coreceptor usage could in principle be derived by directly correlating genotype to phenotype. Potential benefits of a good understanding of genotype-phenotype relations include the possibility to make suggestions that can be validated via site-directed mutagenesis or biophysical studies. Moreover, the genotype reveals a much more fine-grained picture of how "close" a virus is to changing its coreceptor usage behavior. This is similar to the notion of "genetic barrier" which is commonly used in drug resistance to denote the number of mutations a given strain has to accumulate in order to become resistant to a given drug. An understanding of genotype-phenotype relations also has a very practical effect: experimental assays for determining coreceptor usage are expensive and time-consuming, whereas predictions based on the genotype only rely on sequencing a small part of the viral genome which is a cheap and fast routine task. As mentioned in Section 6.1, experimental work has identified the V3 loop as the main determinant of viral coreceptor usage. Consequently, and due to lack of sufficient data for other regions, all current approaches focus on V3.

Early rules-based approaches include (Fouchier et al., 1992; De Jong et al., 1992; Fouchier et al., 1995; Milich et al., 1997; Xiao et al., 1998). Donaldson et al. (1994) presented one of the earliest approaches to predict genotype from phenotype which was not based on a simple amino acids rule. Here, a line was chosen manually to separate R5 from R5X4/X4 sequences in the two-dimensional space spanned by the positive charge of the V3 loop and the number of amino acid differences from an R5

---

<sup>1</sup>The work reported in this section was performed in collaboration with Richard Harrigan and Andrew Low (British Columbia Center for Excellence in HIV/AIDS) and with Martin Däumer and Rolf Kaiser (University of Cologne).

consensus strain.

Statistical learning methods entered the field starts with the work of (Briggs et al., 2000), who proposed a linear model for the combined prediction of coreceptor usage and cell tropism. More precisely, a value of 1 for the dependent variable was used to denote M-tropic R5 virus, a value of 2 for R5X4- or X4-using dualtropic virus, and a value of 3 for X4-using T-tropic virus (while not being an exhaustive listing of coreceptor usage / cell tropism combinations, this choice covered the data used in this study). The explanatory variables were net V3 charge, positive V3 charge, negative V3 charge, and an indicator variable for the presence of an isoleucine at HXB2 position 326 (V3 position 31). These variables were based on a heuristic assessment of V3 alignments and on published evidence.

Beerenwinkel et al. (2001) used decision trees to predict SI/NSI phenotype from V3 sequence. Initially, the amino acids at individual alignment positions were taken as the explanatory variables. In a second run, the overall V3 charge was incorporated as an additional explanatory variable. Resch et al. (2001) applied neural networks for the prediction of coreceptor usage. The dependent variable was coreceptor usage, with one class for R5-virus and one class for R5X4/X4-virus (thus the goal was to detect the capability of using CXCR4). The network topology was specified as a fully connected feed-forward network with 16 input nodes, a hidden layer of three sigmoidal nodes, and one linear output unit. The 16 variables used for the input nodes were the amino acids at positions 5, 7, 8, 10, 11, 13, 18-22, 24, 25, 27, 32, plus the overall charge of the V3 loop. These positions were chosen empirically, guided by some statistical analysis. Amino acids were encoded by numbers from 1 to 21 (including a number for a gap in the alignment), while trying to order the amino acids in a chemically meaningful way (with positively charged amino acids at one end and negatively charged amino acids at the other end).

Pillai et al. (2003) used decision trees (C4.5 and PART) and support vector machines with linear kernel, again for the detection of CXCR4 using virus. Explanatory variables were the individual aligned amino acids in the case of decision trees, and an indicator representation for SVMs (a binary vector with 21 entries for each position, which are all 0 except for the entry representing the amino acid at hand, which is set to 1). This is the first paper which reports a statistically significant overrepresentation of misclassified R5X4 viruses, indicating special problems with this intermediate class.

Jensen et al. (2003) proposed the use of position-specific scoring matrices (PSSMs), for the detection of CXCR4-using virus. PSSMs are a probabilistic sequence model. Sequence profiles (matrices indicating the frequency of each amino acid at each position) are estimated for the class of R5 and R5X4/X4 viruses, respectively. The probability of observing a sequence under one of the two models can be computed from the respective profile when assuming independence among the columns in the alignment. The PSSM score corresponds to a likelihood ratio test between the two

models. Explanatory variables are the amino acids at the individual positions.

A new method for statistical learning called mixtures of localized rules was proposed in (Sing et al., 2004). This approach was used to detect X4-capable virus, based on individual amino acids at the alignment positions. Briefly, simple classification rules are learnt from data in this approach. These rules are combined for prediction in a weighted voting model, where each rule is assigned an individual weight. The weights are estimated by maximizing the area under the ROC curve. Moreover, the weight of each rule can vary across different samples, based on the assumption that a rule's prediction will be most reliable on sequences that are similar to those that were used to learn the rule. Sequence similarity is measured by a sequence profile score, similar to the profiles in (Jensen et al., 2003).

### Performance evaluations

Briggs et al. (2000) used 43 subtype B sequences to fit a linear model. Cutoffs of 1.5 and 2.5 were chosen to transform the numeric predictions into predictions for three classes. They report a perfect model fit (error rate 0%) on the training data and an error rate of 8.33% on an independent test set of 24 isolates with known coreceptor usage.

Beerenwinkel et al. (2001) used 189 sequences, of various subtypes, from 156 patients. Model fit was estimated by 10-fold cross-validation. For decision trees with amino acid input, an average error rate of 10.1% was reported on training sets, and an average of 17.5% on test sets. Adding overall V3 charge as another explanatory variable resulted in an average error rate of 11.6% on both the training and test sets. This compared to error rates of 15.8% and 16.4% on training and test sets, respectively, when refitting the model of Donaldson et al. (1994). Thus, decision trees using amino acids and overall charge represented a significant increase in accuracy as compared to the Donaldson model.

Resch et al. (2001) used 216 subtype B sequences (cloned isolates; 168 R5, 28 R5X4, 17 X4) from 177 patients from the HIV Sequence Database in Los Alamos and from published studies. Patients were carefully shown not to be epidemiologically related. Several neural networks were trained on subsets of the whole data set. The network with the best accuracy on the test set was used to report estimates. Those were compared with the performance of the "charge rule". On the training set, neural network sensitivity, specificity, and phi coefficient were reported as 0.8, 0.98, and 0.82, respectively. Corresponding values for the charge rule were estimated as 0.60, 0.89, and 0.51. Sensitivity, specificity, and phi coefficient on an independent test set were reported as 0.80, 0.89, and 0.67 for the neural network and 0.53, 0.87, and 0.42 for the charge rule.

Pillai et al. (2003) used 271 sequences from multiple subtypes (168 R5, 103 X4, 21 R5X4) from the HIV Sequence Database. Only sequences of length 34 to 36 were included. The database was compiled so that no duplicate sequences were contained in

it. [Jensen et al. \(2003\)](#) used the data set compiled by [Resch et al. \(2001\)](#) to estimate a PSSM. Results were compared to the charge rule and a modified version of the charge rule. In 100 times 10-fold cross-validation, the method was shown to have sensitivity of 0.64 and specificity of 0.94, when using all data. When using unique sequences, the sensitivity dropped to 0.59 and the specificity to 0.86. For comparison, values for the charge rule and for the modified charge rule were reported: the charge rule sensitivity and specificity were 0.58 and 0.90, respectively, and 0.51 (sensitivity) and 0.96 (specificity) for the modified charge rule.

## 6.2.2 Method comparison on clonal data

Here, we provide a comparison of the methods introduced above in a joint cross-validation setting, using a large clonal dataset of HIV V3 genotype-phenotype pairs. The raw output of all methods is a numeric score, optimized for class prediction (R5-only versus X4-capable) via a pre-defined and method-dependent cutoff.

Specifically, the predictive performance of the 11/25 rule is compared with alternatives based on statistical learning ([Resch et al., 2001](#); [Pillai et al., 2003](#); [Jensen et al., 2003](#); [Sing, 2004](#)) using ten replicates of 10-fold cross-validation. The comparison is based on 1,110 clonal genotype-phenotype pairs obtained from the Los Alamos HIV Sequence Database and from selected publications. The samples originate from 332 patients, with 769 R5, 131 R5/X4, and 210 X4 phenotypes.

Clonal and clinical samples were aligned with the multiple alignment package MUSCLE ([Edgar, 2004](#)), using standard parameters, followed by visual inspection. No manual alignment correction was necessary.

R5/X4 and X4 variants were pooled into a single class (X4-capable), as opposed to variants that are limited to using CCR5 (R5-only). 156 samples (14%) had insertions or deletions relative to the subtype B V3 consensus sequence,

CTRPNNNTRKSIHIGPGRAFYTGTGEIIGDIRQAHC,

the reference for position numbering in this study. It should be noted that the V3 region of the HXB2 ([Korber et al., 1998](#)) sequence has two insertions and one deletion relative to this consensus sequence.

The 11/25 rule has a mean sensitivity of 59.5% in detecting X4-capable variants and a mean specificity of 92.5% on the clonal isolates, in good agreement with previous studies. We compared this with a variety of other prediction methods in the framework of receiver operating characteristic (ROC) analysis ([Sing et al., 2005](#)) to analyze the sensitivity/specificity trade-off across the range of all possible cutoffs ([Figure 6.1](#)). These analyses, and all previous studies, are based on genetically and phenotypically homogeneous clonal samples. The ROC curve shows the trade-off between sensitivity and specificity by varying the score cutoff for all compared predictive methods. In

our analysis, decision trees, neural networks, and mixtures of localized rules only led to minor improvements in predictive performance over the 11/25 rule when a method-specific cutoff corresponding to the 11/25 specificity of 92.5% was chosen. In contrast, classifiers based on position-specific scoring matrices (PSSMs) or support vector machines (SVMs) significantly outperformed the 11/25 rule, increasing sensitivity by 12.4 percentage points and 16.9 percentage points, respectively, at the 11/25 rule specificity. Using Wilcoxon's ranks sum test, the differences in sensitivity between the SVM and PSSM ( $p = 0.03$ ) or 11/25 ( $p < 10^{-12}$ ) were significant at this specificity. SVMs and PSSMs also showed significantly higher areas under the ROC curve (AUC) than the three other methods (0.91 and 0.90 respectively), indicating an overall improvement in the ability to distinguish X4-capable samples from R5-only samples.

### 6.2.3 Method comparison on population-based data

In contrast to the scenario reflected by clonal data, it is important to note that the viral population *in vivo* is a swarm of genetically and phenotypically heterogeneous variants, often termed a "quasispecies" (Domingo et al., 1995); therefore, approaches which give satisfactory results on clonal data may not be satisfactory on clinically derived data. In order to obtain a representative sample of this quasispecies, a substantial number of clones would have to be phenotyped and/or genotyped, which is not presently feasible in routine clinical practice. Instead, both the genotype and the phenotype are obtained using bulk or "population-based" approaches. Genotypes from population-based sequencing often contain mixtures of co-existing viral variants. Sequence ambiguities that remain undetected either by genotyping or phenotyping may differ from the exact genotype-phenotype match seen in clonal samples. In this section, we evaluate the implications of using population-based – as opposed to clonal – sampling strategies on the reliability of coreceptor usage prediction.

We analyzed plasma samples from 952 antiretroviral-naive patient samples with matched V3 genotype and coreceptor phenotype. V3 genotype was determined using population-based "bulk" sequencing techniques and coreceptor phenotype was obtained using the Trofile Coreceptor Assay (Monogram Biosciences), as described in (Brumme et al., 2005a). Only one of these samples was phenotyped as pure X4 with the coreceptor assay, suggesting that *in vivo*, X4 virus is only rarely present without concordant R5 variants. We use the term "X4-capable" for samples containing either R5/X4 or X4 variants, in contrast to "R5-only" samples. The sequences were aligned together with the clonal sequences from Section 6.2.2 using the multiple alignment package MUSCLE (Edgar, 2004), with standard parameters, followed by visual inspection. No manual alignment changes were made.

For each mutation we first assessed the association with coreceptor usage in univariate analysis, using Fisher's exact test (mutation vs. wild type amino acid; R5-only vs. X4-capable). Correction for multiple testing was performed using the Benjamini-

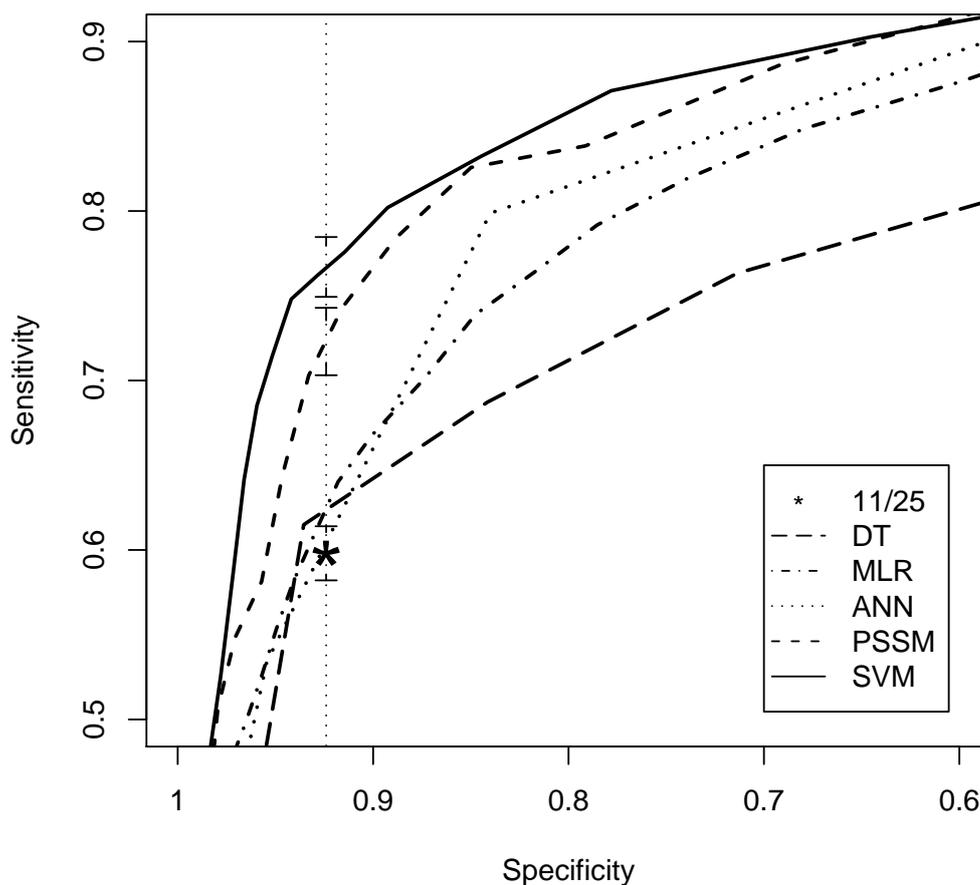


Figure 6.1: Predictive performance of the 11/25 rule and five statistical learning methods, assessed on clonal data. Evaluation was performed using ten replicates of ten-fold cross-validation, followed by threshold averaging of the ROC curves. The dotted vertical line indicates the specificity of the 11/25 rule (92.4%). While decision trees (DT), mixtures of localized rules (MLR), and artificial neural networks (ANN) did not improve substantially over the classical rule in our analysis, PSSMs and SVMs did lead to an increase in sensitivity of 12.4% and 16.9%, respectively. One standard error indicates the sensitivity spread of 11/25 rule, PSSM, and SVM at the specificity of 92.4%.

Hochberg method (Benjamini and Hochberg, 1995), at a false discovery rate of 5%. As shown in Table 6.1, a total of 41 mutations at 19 V3 positions were significantly associated with coreceptor usage (25 predicting X4-capability, 16 predicting R5-only variants).

As mentioned above, we are specifically interested in X4-associated mutations, which occurred at 14 V3 positions. The positions are highlighted in Figure 6.5, on top of the only V3 loop structure in context of the gp120 core available to date (Huang et al., 2005). As can be seen in the figure, most of the significant mutations are located close to the tip of the loop, where the two strands of the  $\beta$ -hairpin are in close spatial proximity. The most well-known V3 mutation 11R ( $p = 1.58 \cdot 10^{-20}$ ) was present in 31 samples (often in mixtures also containing glycine or serine), 29 of which (94%) were phenotyped as X4-capable. However, other mutations were also strong X4 determinants. Mutation 13Y ( $p < 10^{-8}$ ) occurred in 21 samples, 16 of which were phenotyped as X4-capable. In all cases in which samples with 11R or 13Y were phenotyped as R5, sequence ambiguities were present at these positions, possibly indicating cases of genotype-phenotype mismatch. To assess the evolutionary role of 13Y and other mutations, we analyzed mutation covariation between all pairs of the 25 X4-associated V3 mutations, again using Fisher's exact test with the Benjamini-Hochberg method at a false discovery rate of 0.05 (the method is described in detail in Section 5.2.3). As shown in Figure 6.6, mutation 11R occurred as part of a mutation cluster with significant pair-wise association, containing also mutations 9S, 13S, and 24R. Some of these associations had been described in a previous covariation analysis based on mutual information rather than Fisher's exact test (Korber et al., 1993). In contrast, mutation 13Y was not significantly associated with any of these mutations, suggesting the existence of several alternative mutational pathways for evolution from R5-only to X4-capable genotypes.

After this univariate analysis, we evaluated the predictive performance of multivariable models on the population-based data, in comparison to their performance on the clonal data. We focussed on the 11/25 rule (as the classical approach) and the SVM (as the best-performing prediction method on the clonal data; cf. Section 6.2.2). The specificity of the 11/25 rule remained close to that observed for the clonal data (93.5%), while the sensitivity decreased drastically, to 30.5% compared to the 59.9% sensitivity observed on the clonal dataset. SVM-based prediction was evaluated using ten replicates of 10-fold cross-validation. Averaged over the 100 test sets from the cross-validation data, the mean sensitivity of the 11/25 rule was even lower (25.9%) than on the dataset as a whole, at a mean specificity of 93.9%. SVM-based prediction again outperformed the 11/25 rule, but, as compared to the clonal data, also decreased substantially in sensitivity to 39.8% at the 11/25 specificity of 93.5%.

In summary, training and testing SVMs on our clinically derived, population-based genotype-phenotype data set reveals a substantial decrease in sensitivity, as compared

to clonal data reported in Section 6.2.2. The extent of this decrease contrasts with previous experience with the prediction of phenotypic resistance to protease or reverse transcriptase inhibitors. In that case, satisfactory results can be obtained even when using population-based data, and from as few as 500 genotype-phenotype pairs (Beerenwinkel et al., 2002).

We hypothesize that the transition from clonal to clinically-derived data has a more profound impact on predicting coreceptor usage than on predicting phenotypic resistance due to the different nature of the reported phenotype. In drug resistance, the phenotype is reported as a continuous quantity – the fold-change in 50% inhibitory concentration as compared to a reference strain (cf. Section 2.5) – and the presence of mixed populations will only cause the reported phenotype to be more variable around the mean value in repeated experiments, as compared to measurements on homogeneous populations. In contrast, coreceptor usage phenotype is a categorical quantity, and instead of only increasing measurement variability, mixed populations can lead to a complete mismatch between genotype and phenotype. Furthermore, X4 virus is usually present as a mixture or a minority species whereas resistance mutations are often the predominant viral species. The mismatch can be due to X4-capable variants detected in the phenotype, but undetected in the genotype, either because they are a minority species or due to mutations outside the sequencing range. As an example, some of the population-based samples share identical nucleotide sequences within the sequencing range, but are associated with different phenotypes. On the other hand, X4-capable variants may remain undetected in the phenotype assay: in a separate study, we tested 74 of these clinic samples with a second recombinant phenotype assay, finding an assay agreement of only 85.1% (Skrabal et al., 2007).

The most important limitation of the present and previous studies is the exclusive consideration of the V3 region. There is accumulating evidence that other parts of the “bridging sheet” that connects inner and outer domain of gp120 are critically involved in coreceptor selectivity (Pastore et al., 2006) and using key mutations in these regions could substantially increase the sensitivity of any genotype-based prediction algorithm.

Most importantly, by assembling a large dataset on coreceptor usage for population-based samples, we have demonstrated that performance evaluations based on clonal data – the basis for all prediction studies to date – provide positively biased estimates of the predictive reliability of genotypic methods for predicting HIV coreceptor use in clinical practice. In the latter scenario, the prediction problem is complicated by the use of population-based genotyping and phenotyping, and the presence of genetically and phenotypically heterogeneous – possibly even undetected – viral subpopulations. A multitude of V3 mutations appear to be associated with coreceptor usage, in addition to the classical mutations at positions 11 and 25. The role of these mutations should be confirmed *in vitro* using mutagenesis studies in a variety of genetic backgrounds. Our results also show consistent advantages to statistical learning

methods over the classical 11/25 rule.

### 6.3 Clinical and host markers in predicting coreceptor usage<sup>2</sup>

In the previous section we have shown that predicting coreceptor usage from clinically derived (population-based) data is substantially harder than for clonal data. In the present section, we study if the incorporation of additional data only available in the clinical setting can countervail against the performance loss incurred due to population-based (as opposed to clonal) sequencing and phenotyping.

To identify potentially useful markers of coreceptor usage, we examined clinical parameters, including plasma viral load (VL), CD4 and CD8 cell counts, and the percentage of CD4<sup>+</sup> T-cells (CD4%) at the time of sampling for all of the 976 HOMER patients (Brumme et al., 2005a). Moreover, patients were tested for heterozygosity at the CCR5  $\Delta$ 32 allele, a 32-basepair deletion resulting in non-functional CCR5 coreceptors (Brumme et al., 2005b). The univariate association between clinical features and coreceptor usage is summarized in Table 6.1. To assess the predictive benefit of these features when combined with a purely sequence-based prediction approach, we evaluated different feature subsets in combination with an SVM-based classifier. The best feature combination relied on four additional features ( $\log_{10}(\text{CD4}\%)$ ; host CCR5  $\Delta$ 32 heterozygosity; number of ambiguous amino acid V3 positions; and a variable indicating the presence of insertions or deletions in the V3 sequence). The improvements in sensitivity over the 11/25 rule and the purely sequence-based SVM were substantial when the clinical parameters were considered (Figure 6.2). The sensitivity of 63% at the 11/25 specificity corresponds to a 2.4-fold improvement in detecting X4-capable samples relative to the 11/25 rule. SVM-based feature ranking showed that 221 of the 704 variables had non-zero weights in the combined model. In particular, 98 variables contributed to increased CXCR4 propensity, and 123 variables to increased CCR5 propensity (Figure 6.3). The top 5-ranking variables were CD4%, the presence of mutation 13Y, the presence of mutation 11R, the number of ambiguous V3 positions, and the presence of mutation 24G. As shown in Figure 6.4, the probabilistic SVM output was well-calibrated (e.g. approximately 75% of the samples with X4 probability of 75% were indeed X4-capable).

Thus, the intricacies of population-based data can be compensated for by leveraging additional information which is routinely available in the clinical setting, but not with clonal data. Indeed, we have shown that the very source of these intricacies, namely the genetic diversity of the viral population as measured by the number of

---

<sup>2</sup>The work reported in this section was performed in collaboration with Richard Harrigan and Andrew Low (British Columbia Center for Excellence in HIV/AIDS).

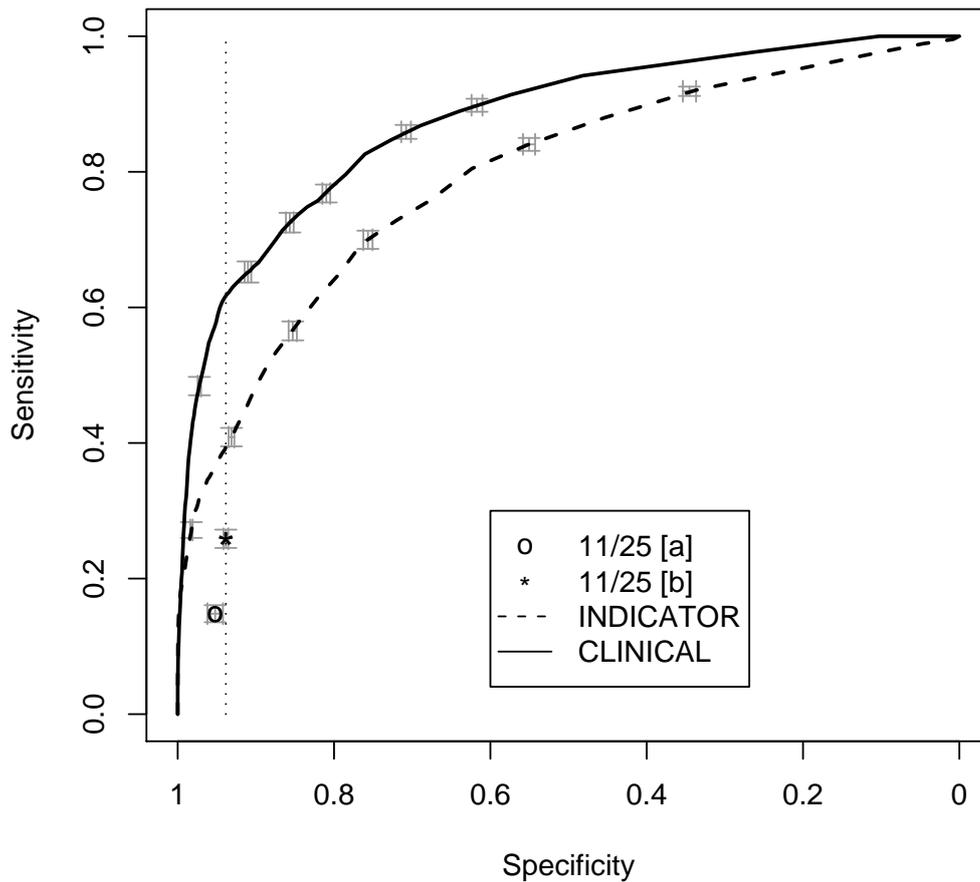


Figure 6.2: Predicting coreceptor usage on population-based data. ROC curves are obtained by threshold averaging from ten replicates of ten-fold cross-validation. Bars indicate one standard error in horizontal and vertical direction. The 11/25 rule performs better when it is applied also in the presence of ambiguous positions 11 or 25 (11/25[b]), as compared to the requirement of unambiguous positions in these places (11/25[a]). Compared to the clonal dataset, a substantial performance decrease is observed not only for the 11/25 rule, but also for the SVM with amino acid indicator representation (SVM). The inclusion of additional features (CD4%, number of sequence ambiguities, host CCR5  $\Delta$ 32 heterozygosity, presence of insertions/deletions), leads to considerable improvements in predictive performance (SVM and CLINICAL).

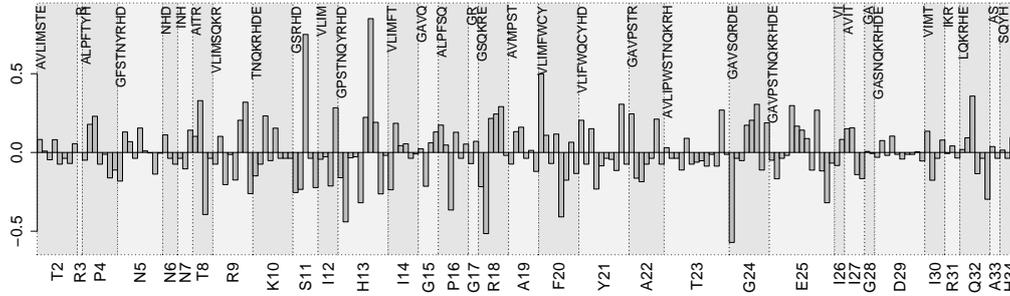


Figure 6.3: Feature ranking for the SVM trained on the clinical dataset (cf. Section 5.2.4 for details). The feature set consists of 704 variables: for each of the 35 V3 positions, 20 variables indicate the presence of specific amino acids. It can be seen that mutation 13Y contributes most strongly to X4 predictions, followed by the 11R mutation. Four additional variables representing CD4%, number of sequence ambiguities, host CCR5  $\Delta 32$  heterozygosity, and presence of insertions/deletions, respectively, are not shown here, but mentioned in the text.

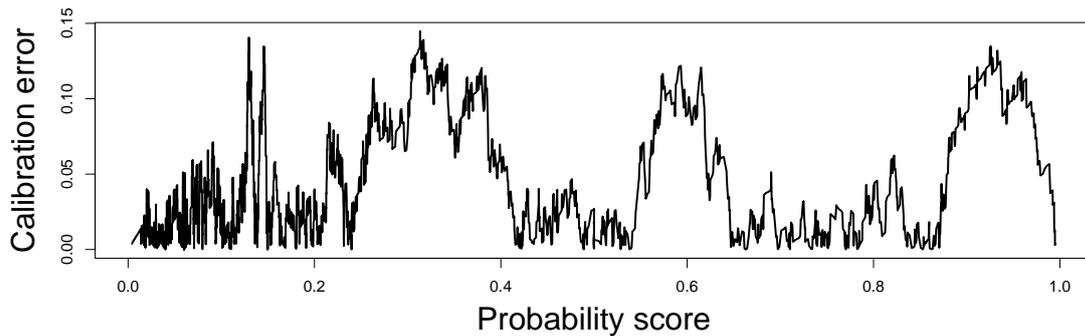


Figure 6.4: Calibration analysis of the probabilistic predictions. The horizontal axis indicates the predicted probability. Ideally, for example, 75% of predictions with confidence 75% should be correct. The calibration error at a given probability is the absolute difference between the two quantities.

Table 6.1: V3 mutations significantly associated with coreceptor usage in univariate analyses of the population-based data. The header row shows the most frequent residue at the V3 position with relative frequency/entropy below. Underlined mutations predict X4-capable virus, while the remaining predict R5-only virus, with relative frequency of *all* (R5-only; X4) samples below. The  $p$ -value (Fisher's exact test) is indicated in brackets as  $\log_{10}$ .

N6	N7	T8	R9	S11
<u>N</u> (-3.8) 97.0 (98.1;92.1)	<u>N</u> (-6.0) 96.7 (98.2;89.9)	<u>I</u> (-4.0) 1.4 (0.4;4.5)	<u>R</u> (-7.9) 93.0 (95.5;82.0)	<u>R</u> (-19.8) 3.2 (0.3;16.3)
	<u>Y</u> (-3.0) 0.4 (0.0;2.2)	<u>T</u> (-3.6) 96.9 (98.0;92.1)	<u>S</u> (-4.2) 2.8 (1.6;7.9)	
			<u>K</u> (-3.7) 0.5 (0.0;2.8)	
I12	H13	R18	A19	F20
<u>I</u> (-3.5) 94.0 (95.4;87.6)	<u>Y</u> (-8.8) 2.3 (0.6;9.6)	<u>S</u> (-2.2) 8.0 (9.3;2.2)	<u>V</u> (-6.0) 3.9 (2.3;11.2)	<u>F</u> (-6.2) 84.4 (87.4;71.3)
	<u>S</u> (-4.6) 4.6 (3.1;11.2)		<u>A</u> (-3.2) 86.5 (88.4;78.1)	<u>V</u> (-5.2) 1.8 (0.8;6.7)
	<u>T</u> (-2.8) 5.0 (3.9;10.1)			
	<u>R</u> (-2.8) 3.8 (2.8;8.4)			
Y21	A22	T23	G24	E25
<u>H</u> (-5.2) 1.8 (0.8;6.7)	<u>A</u> (-4.5) 71.5 (74.5;58.4)	<u>T</u> (-4.6) 93.2 (95.0;85.4)	<u>R</u> (-11.4) 2.3 (0.4;10.7)	<u>D</u> (-4.8) 34.4 (37.4;20.8)
<u>I</u> (-3.8) 0.9 (0.3;3.9)	<u>T</u> (-4.4) 27.0 (24.2;39.9)	<u>A</u> (-2.7) 1.8 (1.1;5.1)	<u>G</u> (-7.9) 84.7 (88.1;69.7)	<u>R</u> (-4.6) 4.6 (3.1;11.2)
		<u>R</u> (-2.7) 1.2 (0.6;3.9)	<u>E</u> (-3.9) 5.7 (4.3;12.4)	<u>Q</u> (-3.8) 7.0 (5.4;14.0)
			<u>S</u> (-3.0) 0.6 (0.1;2.8)	<u>N</u> (-2.6) 2.6 (1.8;6.2)
I26	I27	I30	Q32	
<u>I</u> (-2.6) 92.4 (93.7;86.5)	<u>V</u> (-4.2) 7.1 (5.4;14.6)	<u>I</u> (-2.7) 97.1 (98.0;93.3)	<u>Q</u> (-4.0) 82.9 (85.2;72.5)	
	<u>I</u> (-3.7) 88.8 (90.7;90.3)		<u>K</u> (-3.6) 13.5 (11.5;22.5)	
	<u>A</u> (-3.0) 0.6 (0.1;2.8)			



## 6.3. CLINICAL AND HOST MARKERS

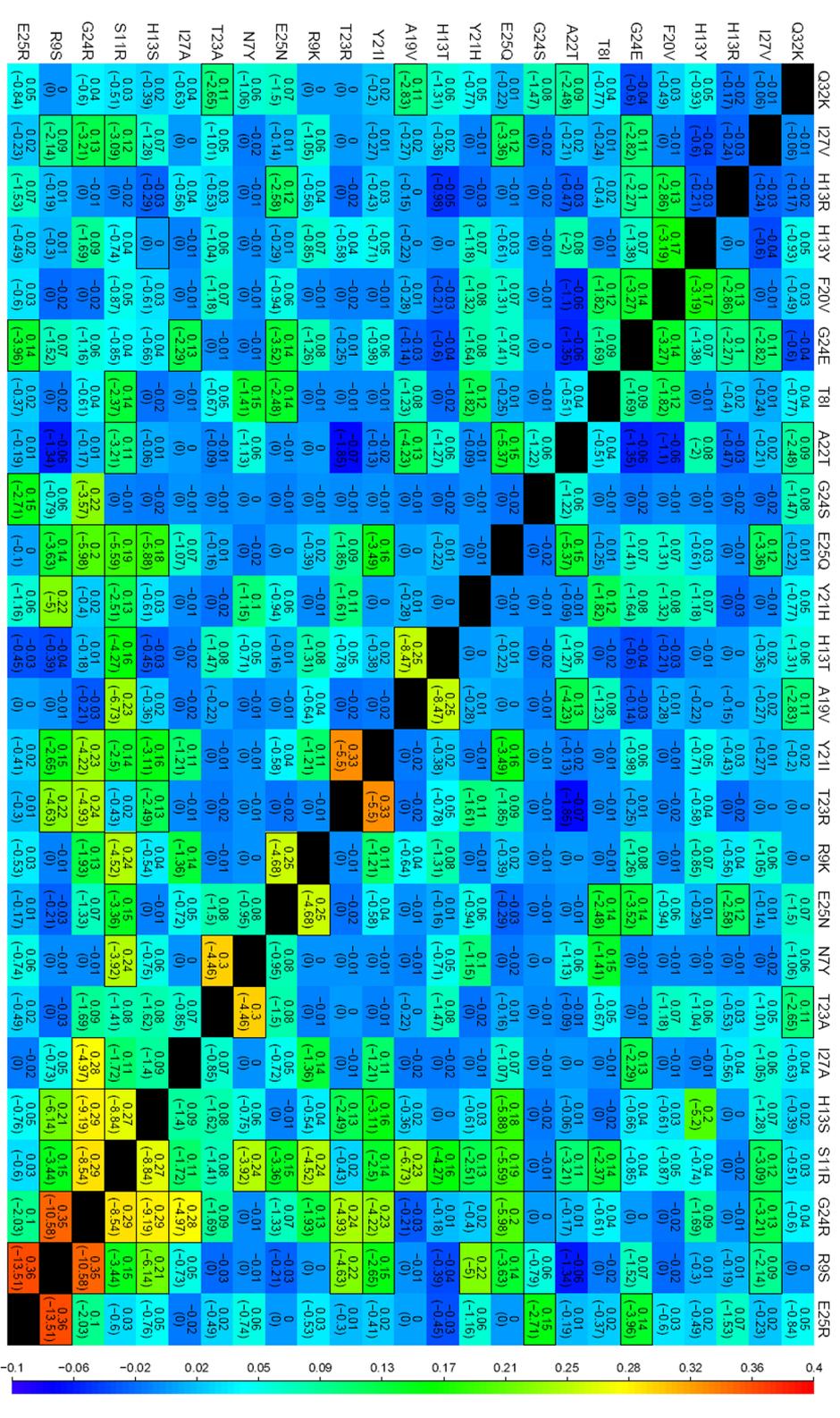


Figure 6.6: Covariation among pairs of X4-associated V3 mutations. Within each box, the upper value indicates the phi correlation coefficient. The lower value (in brackets) shows the  $\log_{10}$  of the p-value from Fisher's exact test. Mutation pairs with covariation significantly from independence after correction for multiple testing are framed with a box. See Section 5.2.3 for details on the analysis method.

ambiguous sequence positions, can be one of the most important predictors of coreceptor usage both in univariate analyses and in SVM-based models. Remarkably, the association between increasing viral genetic diversity and increased CXCR4 propensity observed here on cross-sectional, population-based data supports the model of *env* evolution postulated in the longitudinal study (Shankarappa et al., 1999) on the basis of multiple clones per time point. Beyond the viral genomic information, we show that the host genome (CCR5  $\Delta$ 32 heterozygosity), and more significantly, the host immunological status (as measured by CD4%) contain information relevant for predicting the potential presence of undetected X4-capable variants. While these quantities may or may not be causally related to coreceptor usage, the association with these parameters clearly helps to distinguish environments that are more typical of X4-capable variants. Alternatively, the relationship between the V3-genotype and viral coreceptor phenotype may be dependent on CD4 count or CD4%. These additional data on viral diversity, host factors, and host immunological status contain non-redundant information, as evidenced by the substantial improvements in predictive performance compared to purely sequence-based prediction. Genotype-based prediction could be used exclusively, or as a prefilter to phenotyping (in that phenotyping is only performed when the predicted X4 probability is close to the chosen cutoff, indicating a potentially difficult case).

## 6.4 Structural determinants of coreceptor usage<sup>3</sup>

In this section, we describe a structure-based approach to predicting HIV-1 coreceptor usage, as published in (Sander et al., 2007). The novel approach is shown to be superior to a standard sequence-based method.

### 6.4.1 The structural basis of coreceptor usage

To date, information on the three-dimensional structure of the V3 loop has not been exploited for predicting the coreceptor type used by a viral population. Including structural information can improve predictive performance and, even more importantly, be a first step towards a deeper understanding of the structural aspects of coreceptor usage. Several studies have analyzed conformational properties of the V3 loop. However, these investigations did not particularly consider the impact on coreceptor usage. As Lusso (2006) points out, structural understanding of coreceptor specificity is limited at the moment. In recent work, Watabe et al. (2006) suggested empirical potentials to assess the fit of sequence variants to loop candidates generated by Monte Carlo variation of NMR peptide structures. So far, structural studies have been based on peptide struc-

---

<sup>3</sup>The work reported in this section was performed in collaboration with Oliver Sander and Francisco Domingues (MPI for Informatics, Saarbrücken).

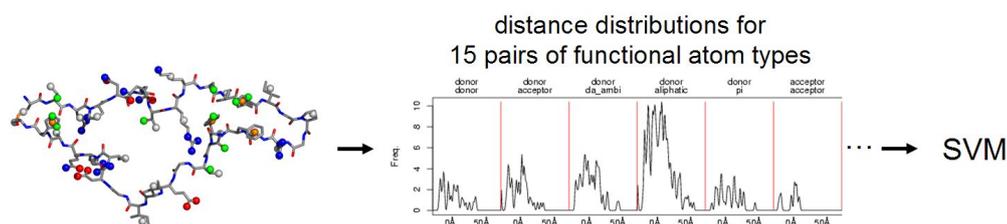


Figure 6.7: Schematic overview of the structural descriptor computation and coreceptor usage prediction.

tures, as no completely resolved structure of gp120 was available. The situation has changed with a recently published crystal structure of the HIV-1 JR-FL gp120 protein (PDB 2b4c) including the V3 loop (Figure 2.4(c)) by Huang et al. (2005). Although some evidence for conformational changes in the loop structure exists, there is an ongoing debate about the relevance of V3 loop conformation to coreceptor selectivity (Sharon et al., 2003; Rosen et al., 2006). Sharon et al. (2003) suggest that alternative conformations of the V3 loop play a key role in determining the coreceptor specificity of HIV-1. On the other hand, Scheib et al. (2006) argue that there is a predominant conformation for both R5 and X4 variants and that varying sequence features are responsible for specificity towards the respective coreceptor.

## 6.4.2 Methods

Here, we propose an approach for incorporating a structural component into our model of coreceptor usage. Structural approaches to predicting coreceptor usage have to face two major challenges: Firstly, structural details on gp120-coreceptor complexation are unknown. This is why we refrain from attempting to integrate structural information on the coreceptors. Secondly, no crystal structures are available for viral variants. As it is unlikely that comprehensive structural data on the wealth of viral variants will become available, modelling of side chains and potentially also changes in the backbone is necessary. The decision to rely on sequence data alone for user input is also justified by application scenarios in routine clinical practice, where structure data will never be feasible to obtain. Thus, the novel component will try to predict the structural changes in the V3 loop induced by sequence alterations. In turn, features will be extracted from the predicted structure which can be used as input for statistical learning methods. In our structural modeling, we rely on a structural template of the V3 loop provided by the single available crystal structure of the V3 loop in context with gp120 (Huang et al., 2005, PDB identifier 2b4c). This structure represents a CCR5-using variant.

We focus on two simple and computationally cheap approaches to structure modeling (Figure 6.7 shows a schematic overview). In both approaches, the V3 loop back-

bone is kept rigid (i.e. invariant under mutations). The first approach, called  $V3SD_{C\beta}$ , approximates the position of all functional side chain atoms by the fixed  $C\beta$  positions of the structure 2b4c. The second approach, called  $V3SD_{SCWRL}$ , models side chains using SCWRL (Canutescu et al., 2003). SCWRL is a reliable and fast program to predict side chains for large sets of sequences. By comparing the two descriptors induced by  $V3SD_{C\beta}$  and  $V3SD_{SCWRL}$  which represent structures of viral variants at two different levels of approximation, the trade-off between increased uncertainty and the improved information about side chain location and length can be assessed. To specifically address the structural uncertainty in the presence of insertions and deletions, we evaluate the performance separately for sequence variants with substitutions only, as opposed to variants exhibiting also insertions and deletions relative to the reference V3 loop of the structure 2b4c.

In order to extract features from the modelled variants (which are often called “structural descriptors” in structure-based bioinformatics), the side chains are represented by functional atoms, labelled as hydrogen-bond donor, acceptor, ambivalent donor/acceptor, aliphatic, or aromatic ring, according to Schmitt et al. (2002). The spatial arrangement of side chains induced by the modeling approaches  $V3SD_{C\beta}$  or  $V3SD_{SCWRL}$  is then encoded by 15 distance distributions, one for each pair of functional atom types. Specifically, for each atom type combination (e.g. donor-donor, donor-acceptor, etc.) all Euclidean distances between the respective atoms are computed and condensed into a distribution function, similar to a smoothed histogram. Finally, the 15 histograms are sampled at fixed positions, and the results concatenated into a single vector. This vector is then taken as the structural descriptor for a given sample, and used as input for support vector classification.

### Technical details

**V3 loop structural template.** In the 2b4c crystal structure, we extracted the V3 loop from chain G, ranging from residue 296 to residue 331.

**The five functional atom types.** Assignment of amino acids to five (overlapping) functional atom types was performed as follows:

- *Hydrogen-bond donor*: R, N, Q, K, W.
- *Hydrogen-bond acceptors*: N, D, Q, E.
- *Ambivalent donor/acceptors*: H, S, T, Y.
- *Aliphatic amino acids*: A, R, C, I, L, K, M, P, T, V.
- *Pi-stacking centers*: H, F, W, Y.

This assignment follows [Schmitt et al. \(2002\)](#), but does not assign backbone centers as pi-stacking. For aliphatic and aromatic interaction centers all involved atom positions were averaged per residue to compute a pseudo-atom. In contrast to [Schmitt et al. \(2002\)](#) who weight atoms by their solvent access for computing the pseudo-atom of aliphatic side chains, we used the unweighted average of the respective carbons as the solvent exposure of the V3 loop can be seen as rather uniform.

**Feature extraction.** For each of the 15 combinations of functional atom types (i.e. donor-donor, donor-acceptor, etc.), pairwise Euclidean distances between the respective pseudo-atoms in the V3 loop are calculated. The number of these distances depends on the number of pseudo-atoms in the two groups. From these distance matrices we derive distance distributions using a kernel density estimate with Gaussian kernel and bandwidth of 1 Ångstrom. The density estimates are then discretized by uniform sampling at intervals of 0.5 Ångstrom, resulting in a 15 (distance distributions for atom type combinations) times 100 (sample points) dimensional vector. The resulting vector is used as a structural descriptor for a given sample, as an alternative to the purely sequence-based indicator representation, and used as input to the statistical learning method. The bandwidth as well as the sampling intervals for the distance-based descriptors have been set to reasonable values based on empirical observations. To keep computation times feasible they were not optimized systematically.

**Evaluation of predictive performance** To assess the predictive performance of the structure-based descriptors, we compared the two variants  $V3SD_{C\beta}$  and  $V3SD_{SCWRL}$  against purely sequence-based predictions by the 11/25 rule, and by a sequence-based SVM (*Indicator*). The traditional 11/25 rule is an empirically derived procedure often used in clinical practice to predict coreceptor usage. It predicts a viral variant to be X4 if there is a positively charged amino acid at V3 position 11 or 25 ([de Jong et al., 1992](#)). Among simple sequence rules (i.e. not based on statistical learning), [Resch et al. \(2001\)](#) consider the 11/25 rule to be the best predictor of coreceptor usage. The *Indicator* approach is based on a binary sequence encoding. A viral variant is encoded by a bit vector (i.e. consisting of only zeros and ones). Each component in this vector indicates the presence or absence of a specific amino acid at a specific V3 position. In order to evaluate potential benefits of a hybrid approach integrated sequence- and structure-based information, we implemented the  $V3SD_{SCWRL} + Indicator$  representation based on a concatenation of the individual representations.

In order to assess predictive performance for the structural descriptors, we performed ten replicates of ten-fold cross-validation. Evaluation of predictive performance was done using ROCR ([Sing et al., 2005](#)).

Table 6.3: Performance according to various measures for the sequence-based, structure-based, and hybrid approaches. Sensitivity and PPV are given at the specificity of the 11/25 rule, accuracy is measured for the cutoff 0.5.

	11/25 rule	Indicator	V3SD <sub>C<math>\beta</math></sub>	V3SD <sub>SCWRL</sub>	V3SD <sub>SCWRL</sub> + <i>Indicator</i>
Sensitivity	0.6186	0.7340	0.6959	0.7742	0.8041
PPV	0.7692	0.7980	0.7894	0.8065	0.8124
Accuracy	0.8727	0.9000	0.8933	0.9126	0.9156
AUC	0.7824	0.9215	0.9122	0.9266	0.9348

**Support vector classification.** For the sequence indicator encoding (*Indicator*) a linear kernel is used, as previous studies showed that non-linear kernels do not help for simple sequence encodings (Sing et al., 2004). For prediction based on the structural descriptors a radial basis function kernel (Borges, 1998) is applied, as it provides better performance than a linear kernel in this case. In both cases, probabilistic predictions are obtained from the SVM by the method of Platt (1999) to obtain estimates of prediction confidence and a scoring classifier for the receiver operating characteristic (ROC) analysis. To optimize SVM parameters we conducted parameter grid searches. For the linear kernel (*Indicator*) we varied the cost parameter  $\log_2(C)$  in  $[-7, 2]$ . For the radial kernel (V3SD<sub>C $\beta$</sub> , V3SD<sub>SCWRL</sub>) we varied the cost parameter  $\log_2(C)$  in  $[-6, 5]$  and the  $\gamma$  parameter  $\log_2(\gamma)$  in  $[-15, -5]$ . Optimal parameter values were obtained from 10 bootstrap samples of the data set and kept fixed for the subsequent analysis and evaluation. Each bootstrap sample contained 9/10 of the number of samples in the original data set (drawn with replacement), using the default in the R package *e1071*.

**Data set and sequence alignment** We evaluated the two structural descriptors and the two sequence-based predictors on data compiled from the Los Alamos HIV Sequence Database and several publications (Resch et al., 2001; Cilliers et al., 2003; Johnston et al., 2003; Zhang et al., 2002; Zhong et al., 2003). The evaluation was performed on a data set containing 514 mutually distinct V3 sequences. Each of the sequences was annotated as either using CCR5 only or being capable of using CXCR4.

As measures of performance we used the sensitivity at the specificity of the 11/25 rule, the area under the ROC curve (AUC), the accuracy at a cutoff of 0.5 (for the posterior probability obtained by the SVM), and the positive predictive value (PPV) at the specificity of the 11/25 rule. Of all these measures we consider the sensitivity at the specificity of the 11/25 rule as most important in practice, because it focuses on detecting X4 viral variants at an acceptable level of false positives (R5 erroneously considered to be X4).

### 6.4.3 Results

Figure 6.8 shows ROC curves for the different approaches. ROC curves plot (1-specificity) against sensitivity for varied decision cutoffs, ranging from predicting mainly R5 (towards the lower left corner) to predicting mainly X4 (towards the upper right corner). On our data set, the 11/25 rule has a sensitivity of 0.6186 while exhibiting a specificity of 0.9463. Considering the routine clinical application of this simple rule, the benefit of improving the sensitivity towards X4 viral variants is obvious. For the fixed specificity of 0.9463 (i.e. maintaining a fixed number of false positives) the sequence-based indicator prediction using a linear SVM improves sensitivity to 0.7340. A similar improvement has been reported previously (Pillai et al., 2003; Sing et al., 2004) when applying statistical learning methods in comparison to the traditional 11/25 rule.

For the simpler form of structural descriptor  $V3SD_{C\beta}$ , the performance is below the *Indicator* prediction at a sensitivity of 0.6959. Still, this constitutes a considerable improvement over the 11/25 rule. Thus, as features different from pure sequence information are encoded in this structural descriptor, its analysis can provide important insights regarding structural features.

Using structural models for the sequence variants with side chains placed by SCWRL (Canutescu et al., 2003), predictive performance improves considerably over the simple structural descriptor  $V3SD_{C\beta}$  and even compared to the *Indicator* encoding. The structural descriptor  $V3SD_{SCWRL}$  improves sensitivity to 0.7742. SCWRL faces a difficult task in optimizing side chain conformations as no direct contacts between the side chains within the loop with side chains of binding partners are present. However, the improved predictive performance indicates that the additional information over the  $V3SD_{C\beta}$  descriptor helps in discriminating coreceptor usage. One important aspect might be the information about side chain length and volume, which is completely lost in the  $V3SD_{C\beta}$  descriptor.

An overview of predictive performance for further performance measures can be found in Table 6.3. The observed ordering of methods regarding performance is similar to the trend observed for the sensitivities. The absolute performance increases regarding AUC and accuracy are smaller. This is because AUC and accuracy are less responsive to improvements in detection of X4 variants due to the class imbalance towards R5 samples. The relative improvements in sensitivity were significant, both from 11/25 to *Indicator* ( $p = 0.0059$ , paired Wilcoxon test), and from *Indicator* to  $V3SD_{SCWRL}$  ( $p = 0.0137$ ).

Considering the different type of information used in the sequence-based and in the structural descriptors we combined the respective features to assess whether further predictive improvements are feasible ( $V3SD_{SCWRL} + \textit{Indicator}$ ). The sequence-based and structural features were combined by concatenating the corresponding feature vectors. As shown in Figure 6.8, this combination of the sequence-based *Indicator* encod-

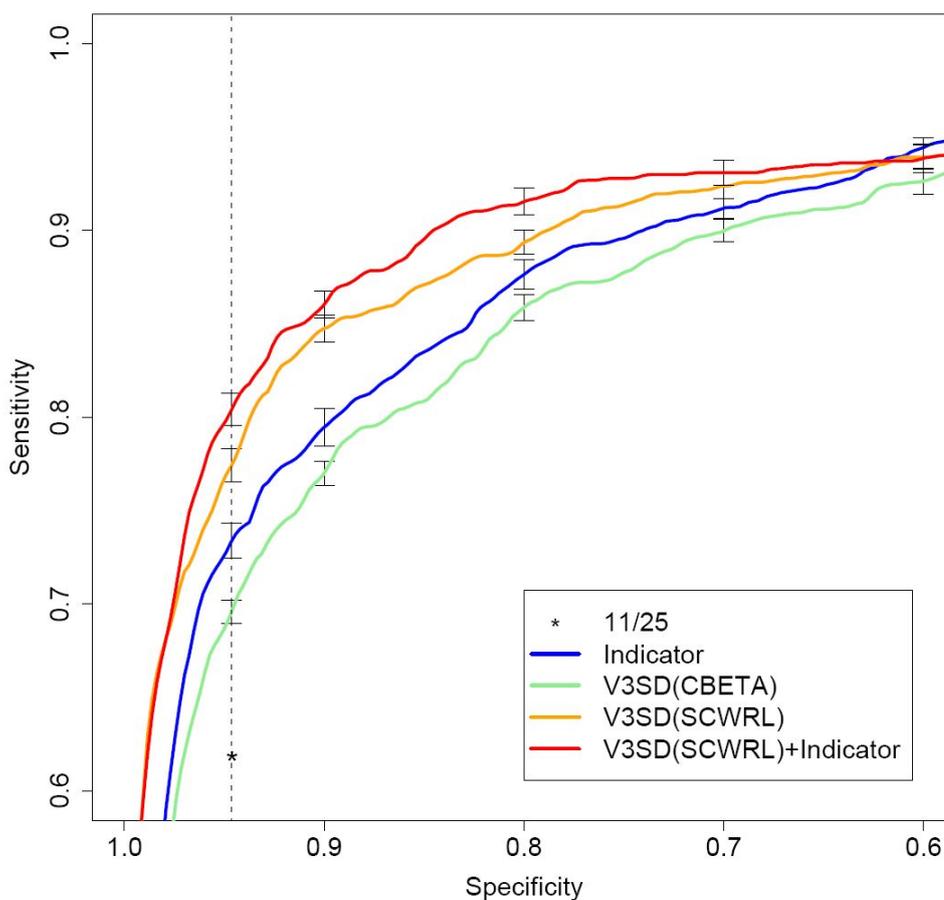


Figure 6.8: ROC comparison of predictive performance: sequence-based (11/25 rule; *Indicator*), structure-based ( $V3SD_{C\beta}$ ;  $V3SD_{SCWRL}$ ), and hybrid ( $V3SD_{SCWRL} + Indicator$ ) predictions.

ing and the structural descriptor  $V3SD_{SCWRL}$  further improves sensitivity to 0.8041 at the specificity of the 11/25 rule (0.9463). This indicates that sequence and structure convey complementary information, to some extent. The sensitivity improvements were significant both from  $V3SD_{SCWRL}$  to  $V3SD_{SCWRL} + Indicator$  ( $p = 0.0098$ ), and from  $Indicator$  to  $V3SD_{SCWRL} + Indicator$  ( $p = 0.0020$ ). Further performance evaluations are shown in [Table 6.3](#).

#### 6.4.4 Discussion

The proposed descriptor yields a considerable performance increase over the established 11/25 rule and even compares favorably to other modern methods based on statistical learning (*Indicator*). In contrast to purely sequence-based coreceptor usage predictions the proposed structural representation captures the relative three-dimensional arrangement of chemical groups. From a biophysical perspective this relative placement of chemical groups determines which coreceptor the viral variant will bind to. The method has also been shown to be robust with respect to sequence variants containing indels ([Sander et al., 2007](#)), making it applicable in realistic scenarios and on large-scale data sets. The most interesting aspect of the proposed descriptor is its integration of structural data, providing the first application of structural data in the context of coreceptor usage prediction. The combination of methods from structural bioinformatics with statistical learning methods allows for competitive performance as well as interpretation of coreceptor usage at the structural level (cf. [Sander et al., 2007](#)).

The proposed structure representation is related to ideas from protein structure comparison and prediction. Distributions of atomic distances have been used successfully in structure comparison ([Carugo and Pongor, 2002](#); [Choi et al., 2004](#)). In protein structure prediction, distributions of distances have been applied as knowledge-based potentials to evaluate the fit of a sequence to a specific structure ([Sippl, 1995](#); [Domingues et al., 1999](#)). In the context of protein function, [Stahl et al. \(2000\)](#) have used distance-based descriptions to cluster active sites of enzymes based on chemical and geometric properties. For the analysis of protein-protein interaction interfaces, [Mintseris and Weng \(2003\)](#) have proposed atomic contact vectors which consist of contact counts derived from thresholded distance matrices. [Aloy and Russell \(2002\)](#) have suggested empirical potentials to assess the compatibility of a pair of sequences to the contacts formed in a known complex of two respectively homologous sequences. In a similar setting, MULTIPROSPECTOR ([Lu et al., 2002](#)) uses a threading algorithm to align a pair of sequences to a structurally resolved protein-protein complex. Besides the interface energy term like in ([Aloy and Russell, 2002](#)), this method also uses the threading score for the protomers themselves.

This study has provided proof of concept that integrating structural modeling can improve the predictive performance over state-of-the-art sequence-based approaches. Despite the good performance there are several limitations and possible directions for

improvement, either by methodological enhancements or by integration of further experimental data. As almost no side chain interactions take place within the V3 loop and the binding partner is not available in the structural model, SCWRL faces a difficult task in optimizing side chains. One possible way of relaxing this difficulty is by considering ensembles of alternative side chain conformations in the structural descriptor. From a methodological point of view, alternative conformations are easy to integrate into the distance distributions in a weighted manner. A further possible bottleneck is the assumption of a fixed backbone structure. Further understanding of the structure-function relationship of coreceptor usage or new insights in the debate mentioned above ([Sharon et al., 2003](#); [Scheib et al., 2006](#); [Rosen et al., 2006](#)) could be incorporated into the descriptor. Instead of the fixed backbone structure, several alternatives are possible. Experimentally resolved peptide structures could be used to model sequence variants or molecular dynamics simulations could be used to generate ensembles of backbone variants. With all these alternatives, the proposed descriptor provides a generic way of incorporating new structural information on V3 loop conformation, especially interesting would be crystal structures of X4 viral variants.

Another interesting perspective is to correlate the discriminative spatial features of the V3 region to spatial arrangements in the coreceptor. Published chemokine receptor models ([Paterlini, 2002](#); [Zhang et al., 2006](#)) could be used to generate such spatial descriptions and search for complementary arrangements of physicochemical properties. Finally, the proposed method to describe the spatial arrangement of physicochemical properties is not limited to the demonstrated application, in principle. By providing a vectorial representation of a binding site it can be used as a generic way of describing and comparing any set of binding sites regarding geometric and physicochemical features involved in different protein-protein interactions.

---

# Summary

Modern anti-HIV therapy, in which three or more drugs are given in combination, can substantially delay disease progression, prolong survival and maintain quality of life. Still, a cure for HIV infection remains out of reach. The main obstacle to ultimate treatment success lies in the emergence of drug-resistant variants in response to incomplete suppression of viral replication. Upon therapy failure, the treating physician is faced with selecting an optimal new drug combination. This task is highly complex due to the ever-increasing number of available antiretroviral drugs, significant cross-resistance and the likely presence of archived drug-resistant viral variants selected by previous regimens. Parameters with potential impact on treatment decisions include the plasma viral load, CD4<sup>+</sup> cell count, viral genotype, phenotype, fitness, pharmacological data, and the treatment history of a patient.

To date, physicians have to rely on a mix of intuition, experience, and empirically derived guidelines and resistance interpretation tables for devising a treatment strategy given the data available for a patient. Model-based approaches to anti-HIV therapy aim to establish a quantitative, comprehensive, and data-driven framework for studying the interactions between host, virus and drugs. Such a framework will also provide a basis for urgently needed decision support systems based on objective and reproducible criteria. In this thesis, we present progress on the elements of model-based therapy, covering many different aspects relevant to optimal therapy choice, in the fields of viral evolution, drug resistance, and coreceptor usage.

The ability of viral populations to adapt extremely fast to changing environmental conditions is the fundamental obstacle to prolonged treatment success. Natural selection determines the “direction” of adaptation. In many situations, the effects of selection can be described in terms of an idealized “fitness landscape” which provides a simple scalar-valued summary of the various factors in the life cycle of organisms that affect their reproductive success. Fitness interactions among pairs of genetic loci, quantified by the classical concept of epistasis, have been shown to be related to the evolutionary fate of a population in a variety of scenarios. However, higher-order interactions cannot be adequately captured in this classical framework.

We analyze a recently proposed classification system for fitness landscapes based on their combinatorial-geometric shape that generalizes epistasis to multiple loci. Using large-scale numerical simulations, we provide proof of concept that the shape of a fitness landscape is significantly associated with the evolutionary fate of a population.

Virtually all longitudinal studies of viral evolution focus on the accumulation of resistance-associated mutations under drug pressure, and little is known about the dynamics of mutations in the absence of drug pressure. While treatment interruptions are not a recommended treatment strategy against HIV, unplanned interruptions are an inevitable part of antiretroviral therapy for many patients (e.g. due to non-adherence or suboptimal drug levels). Our analysis, introducing a novel approach to longitudinal genotype data based on survival analysis, strongly suggests that mutation loss does not revert the complex patterns of mutation accumulation observed during antiretroviral therapy. Rather, mutations are shown to disappear largely independently from each other, albeit at individual rates.

It is not straightforward how to incorporate evolutionary modeling into approaches for predicting drug resistance or response to therapy from genotype. A variety of approaches, from explicit search strategies to direct statistical learning have been proposed. We introduce a more principled approach based on kernel methods. Specifically, we derive a Fisher kernel for mixtures of mutagenetic trees, a family of graphical models for describing the accumulation of resistance mutations under drug pressure. The novel kernel quantifies the similarity of evolutionary escape from antiviral drug pressure between two viral sequence samples. We compare the kernel to a standard, evolution-agnostic amino acid encoding in the prediction of HIV drug resistance from genotype, using support vector regression. The results show significant improvements in predictive performance across 17 anti-HIV drugs.

After these studies related to viral evolution, we turn our focus to the relation between viral genotype, drug resistance, and response to therapy. Most of the approaches proposed to date have been applied in a purely “black-box” manner, without attempting to use them for knowledge discovery. We describe several approaches for extracting biological knowledge from viral genotype and phenotype data. Novel mutations in the HIV genome associated with treatment failure were identified by mining a relational clinical database. Cluster analysis and multidimensional scaling were used to identify the association of novel mutations with specific mutational pathways. Feature ranking based on support vector machines indicated a prominent role of novel mutations in determining phenotypic resistance and in resensitization effects.

We then turn our main focus from phenotypic quantities measured *in vitro* to virological response *in vivo*. The appropriate approach for predicting virological response from genotype is highly controversial. The use of phenotype data obtained under controlled laboratory conditions may seem intuitive, but has faced substantial objections. However, we show in a large-scale study that the use of an input representa-

tion based on predicted phenotypes is clearly competitive with (and often even outperforms) genotype-based representations. Moreover, substantial synergies exist between these different representations, which can be explained by the different kinds of data from which they are derived. These synergies can be exploited in “hybrid” systems combining both sources of information. We also studied the use of predicted replication capacity for clinical purposes; however, the observed benefits were marginal.

The bulk of computational work related to anti-HIV therapy has been performed in the context of the classical drug classes of protease and reverse transcriptase inhibitors. However, with drugs from entirely novel classes already approved (fusion inhibitors) or expected to see approval before 2008 (integrase inhibitors, coreceptor antagonists), decision support has to be extended to these classes. Here, we focus on the support of treatment with coreceptor antagonists. After a comprehensive review of the field, we analyze the relation between genotype and coreceptor usage phenotype on both clonal and clinical (“population-based”) data. We exploit clinical and host information as surrogate markers for undetected viral minority populations. Finally, we show how to further improve predictive performance by applying simple and fast molecular modeling methods.

Several of the methods developed in the context of this thesis have been made available to the public, either as packages for the statistical language R (ROCR; covaRius) or in the form of web-based services (mutation scoring for geno2pheno; geno2pheno[coreceptor]).



---

# Zusammenfassung

Moderne Kombinationstherapien gegen HIV, bei welchen drei oder mehr Wirkstoffe in Kombination verabreicht werden, können das Fortschreiten der Erkrankung wesentlich verlangsamen und zu verlängertem Überleben und verbesserter Lebensqualität beitragen. Trotz dieser Fortschritte ist ein Heilen von HIV-Erkrankten derzeit unmöglich. Ein endgültiger Behandlungserfolg wird hauptsächlich verhindert durch das Auftreten wirkstoffresistenter Varianten in Folge von unvollständiger Unterdrückung der viralen Replikation. Beim Versagen einer Therapie muss der behandelnde Arzt eine optimale neue Wirkstoffkombination auswählen. Diese Aufgabe ist hochkomplex, vor allem durch die ständig wachsende Zahl antiviraler Wirkstoffe, signifikante Kreuzresistenzen, und das Vorhandensein archivierter viraler Varianten, die von früherer verabreichten Medikamentenkombinationen selektioniert wurden. Therapieentscheidungen werden beeinflusst durch Viruslast im Plasma, CD4<sup>+</sup>-Zellzahl, viralen Genotyp, Phänotyp, Fitness, pharmakologische Daten, sowie die Behandlungshistorie eines Patienten.

Derzeit müssen sich Ärzte bei Ihren Behandlungsstrategien auf eine Mischung aus Intuition, Erfahrung, sowie auf empirisch abgeleitete Richtlinien und Interpretationstabellen zur Resistenz verlassen. Modellbasierte Zugänge zur Anti-HIV-Therapie zielen hingegen auf eine quantitative, umfassende, und datengetriebene Basis zum Studium der Interaktionen zwischen Wirt, Virus und Wirkstoffen. Solch eine Basis wird auch für dringend benötigte Systeme zur Entscheidungsunterstützung aufgrund objektiver und reproduzierbarer Kriterien den geeigneten Rahmen bilden. In dieser Arbeit berichten wir über Fortschritte in der modellbasierten Therapie hinsichtlich vieler verschiedener Aspekte im Bereich von viraler Evolution, Wirkstoffresistenz und Korezeptorbenutzung, die für eine optimale Therapieauswahl relevant sind.

Die Fähigkeit viraler Populationen, sich extrem schnell wechselnden Umweltbedingungen anzupassen, ist ein Haupthindernis zu nachhaltigen Behandlungserfolgen. Die "Richtung" der Adaptation wird bestimmt durch die natürliche Selektion. Selektionseffekte können häufig mithilfe einer idealisierten "Fitnesslandschaft" beschrieben werden, welche verschiedene Faktoren im Lebenszyklus eines Organismus, die sich auf den reproduktiven Erfolg auswirken, zusammenfasst. Fitnessinteraktionen zwi-

schen Paaren genetischer Loci werden durch das klassische Konzept von Epistasie quantifiziert und wurden in verschiedensten Szenarien mit dem evolutionären Schicksal einer Population in Verbindung gebracht. Interaktionen höherer Ordnung können jedoch in diesem klassischen Rahmen nicht adäquat beschrieben werden. Wir analysieren ein kürzlich vorgeschlagenes Klassifikationssystem für Fitnesslandschaften basierend auf deren kombinatorisch-geometrischer Form, welches Epistasie auf mehr als zwei Loci verallgemeinert. Mittels umfangreicher numerischer Simulationen zeigen wir, dass die Form einer Fitnesslandschaft mit dem evolutionären Schicksal einer Population signifikant assoziiert ist.

Longitudinale Studien zur viralen Evolution haben sich immer auf die Akkumulation resistenz-assoziiierter Mutationen unter Wirkstoffdruck konzentriert, so dass wenig über die Dynamik von Mutationen in der Abwesenheit von Wirkstoffen bekannt ist. Zwar sind Behandlungsunterbrechungen keine empfohlene Therapiestrategie gegen HIV; ungeplante Unterbrechungen gehören jedoch für viele Patienten unvermeidbar zur antiretroviralen Therapie (z.B. durch Nichtadhärenz oder durch suboptimale Wirkstoffkonzentrationen). Unsere Analyse, die einen neuen Zugang zu longitudinalen Genotypdaten beinhaltet, zeigt, dass der Mutationsverlust während der Behandlungsunterbrechung nicht die komplizierten Muster der Mutationsanhäufung während der antiretroviralen Therapie umkehrt. Die Mutationen verschwinden vielmehr unabhängig voneinander, wenn auch mit verschiedenen Raten.

Es ist nicht offensichtlich, wie sich evolutionäre Modellierung am besten in Methoden zur Vorhersage von Wirkstoffresistenz oder Ansprechen auf Therapie integrieren lässt. Eine Reihe von Methoden, von expliziten Suchstrategien bis zu direkter Verwendung statistischer Lernverfahren wurde hierzu vorgeschlagen. Wir führen hier einen fundamentalen Zugang, basierend auf Kernmethoden, ein. Insbesondere leiten wir einen Fisher-Kern für Mixturen mutagenetischer Bäume her. Hierbei handelt es sich um eine Familie grafischer Modelle zur Beschreibung der Akkumulation von Resistenzmutationen unter Wirkstoffdruck. Der neue Kern quantifiziert die Ähnlichkeit der evolutionären Flucht zweier Sequenzen vor antiviralem Medikamentendruck. Wir vergleichen den Kern mit einer evolutions-agnostischen Standardkodierung von Aminosäuren für die Vorhersage von HIV-Wirkstoffresistenz aus dem Genotyp, basierend auf Support-Vektor-Regression. Die Resultate zeigen signifikante Verbesserungen in der prädiktiven Performanz über 17 Anti-HIV-Medikamente hinweg.

Nach diesen Studien zur viralen Evolution wenden wir uns der Beziehung zwischen viralem Genotyp, Wirkstoffresistenz und Ansprechen auf Therapie zu. Die meisten Methoden, die bisher vorgeschlagen wurden, wurden in einer reinen "black-box"-Art angewendet, ohne den Versuch, sie zur Entdeckung neuen Wissens zu gebrauchen. Wir beschreiben mehrere Zugänge zur Extraktion biologischen Wissens aus viralen Genotyp- und Phänotypdaten. Neue Mutationen im HIV-Genom, die mit Therapieversagen assoziiert sind, wurden durch das Durchsuchen einer relationalen klinischen

Datenbank identifiziert. Clusteranalyse und multidimensionale Skalierung wurden verwendet, um die Assoziation der neuen Mutationen mit bestimmten Mutationspfaden zu charakterisieren. Merkmals-Ranking basierend auf Support-Vektor-Maschinen zeigte eine deutlich ausgeprägte Rolle der neuen Mutationen in der Bestimmung von phänotypischer Resistenz, sowie in Resensitivierungseffekten.

Nach diesen Analysen wenden wir uns von phänotypischen Größen, die *in vitro* gemessen werden, hin zum virologischen Ansprechen *in vivo*. Es ist höchst kontrovers, welcher Zugang zur Vorhersage des virologischen Ansprechens aus dem viralen Genotyp heraus am angebrachtesten ist. Die Verwendung phänotypischer Daten, die unter kontrollierten Laborbedingungen erhoben wurden, mag intuitiv naheliegend erscheinen. Diesem Ansatz wurde jedoch starker Widerspruch entgegengebracht. Wir zeigen aber unter Verwendung umfangreicher klinischer Daten, dass die Verwendung einer Eingabe-Repräsentation, die auf vorhergesagten Phänotypen basiert, sich gut mit rein genotyp-basierten Repräsentationen vergleicht, diese häufig sogar in prädiktiver Performanz übertrifft. Darüberhinaus zeigen wir, dass substantielle Synergien zwischen beiden Zugängen bestehen, die sich durch die unterschiedlichen Arten von Daten, die jeweils verwendet werden, erklären lassen. Diese Synergien können in "hybriden" Systemen, die auf beide Informationsquellen zurückgreifen, ausgenutzt werden. Wir haben darüberhinaus auch noch die Verwendung von vorhergesagter Replikationskapazität für klinische Zwecke untersucht, der beobachtete Nutzen war jedoch marginal.

Der Hauptteil an quantitative Studien im Bereich der Anti-HIV-Therapie wurde im Kontext der klassischen Wirkstoffklassen von Inhibitoren der Protease oder der Reversen Transkriptase durchgeführt. Nachdem jedoch Wirkstoffe von völlig neuartigen Klassen entweder schon zugelassen sind (Fusionsinhibitoren) oder höchstwahrscheinlich noch vor 2008 zugelassen werden (Integraseinhibitoren und Korezeptorantagonisten), muss die Entscheidungsunterstützung auf diese Klassen erweitert werden. Wir widmen uns hier der Unterstützung der Behandlung mit Korezeptorantagonisten. Nach einem umfassenden Überblick über das Feld analysieren wir die Beziehung zwischen Genotyp und Korezeptorbenutzungsphänotyp, sowohl auf klonalen als auch auf klinischen ("populations-basierten") Daten. Wir nutzen klinische und Wirtsinformation als Surrogatmarker für unentdeckte virale Minderheitspopulationen. Abschließend zeigen wir, wie sich die prädiktive Performanz durch die Verwendung einfacher und schneller Methoden des molekularen Modellierens noch weiter verbessern lässt.

Einige der Methoden, die im Rahmen dieser Arbeit entwickelt wurden, sind in öffentlich zugänglichen Implementationen verfügbar, entweder in Form von Paketen für die Statistiksprache R (ROCR; covaRius) oder in Form von Web-basierten Angeboten (Mutations-Scoring für geno2pheno; geno2pheno[coreceptor]).



---

# Bibliography

- Abrahams, E., Ginsburg, G. S., and Silver, M. (2005). The Personalized Medicine Coalition: goals and strategies. *Am J Pharmacogenomics*, 5(6):345–355.
- Alkhatib, G., Combadiere, C., Broder, C. C., Feng, Y., Kennedy, P. E., Murphy, P. M., and Berger, E. A. (1996). CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. *Science*, 272(5270):1955–1958.
- Allman, E. and Rhodes, J. (2004). *Mathematical Models in Biology*. Cambridge University Press.
- Aloy, P. and Russell, R. B. (2002). Interrogating protein interaction networks through structural biology. *Proc Natl Acad Sci U S A*, 99(9):5896–5901.
- Altmann, A., Beerenwinkel, N., Sing, T., Savenkov, I., Däumer, M., Kaiser, R., Rhee, S.-Y., Fessel, W. J., Shafer, R. W., and Lengauer, T. (2007). Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antiviral Therapy*, 12(2):169–178.
- Amari, S. and Nagaoka, H. (2000). *Methods of information geometry*. American Mathematical Society.
- Annan, K. A. (2000). The role of the united nations in the 21st century. Technical report, United Nations.
- Bachelor, L., Winters, B., and Nauwelaers, D. (2004). Estimation of phenotypic clinical cutoffs for virtualphenotype through meta analyses of clinical trial and cohort data. *Antivir Ther*, 9:S154.
- Baker, D. and Sali, A. (2001). Protein structure prediction and structural genomics. *Science*, 294(5540):93–96.

- Balduin, M., Sierra, S., Däumer, M. P., Rockstroh, J. K., Oette, M., Fätkenheuer, G., Kupfer, B., Beerenwinkel, N., Hoffmann, D., Selbig, J., Pfister, H. J., and Kaiser, R. (2005). Evolution of hiv resistance during treatment interruption in experienced patients and after restarting a new therapy. *J Clin Virol*, 34(4):277–287.
- Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature*, 226(5252):1209–1211.
- Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):868–871.
- Beerenwinkel, N. (2004). *Computational Analysis of HIV Drug Resistance Data*. PhD thesis, Saarland University.
- Beerenwinkel, N., Däumer, M., Oette, M., Korn, K., Hoffmann, D., Kaiser, R., Lengauer, T., Selbig, J., and Walter, H. (2003a). Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res*, 31(13):3850–5.
- Beerenwinkel, N., Däumer, M., Sing, T., Rahnenführer, J., Lengauer, T., Selbig, J., Hoffmann, D., and Kaiser, R. (2005a). Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *J Infect Dis*, 191(11):1953–1960.
- Beerenwinkel, N. and Drton, M. (2007). A mutagenetic tree hidden Markov model for longitudinal clonal HIV sequence data. *Biostatistics*, 8(1):51–71.
- Beerenwinkel, N., Lengauer, T., Däumer, M., Kaiser, R., Walter, H., Korn, K., Hoffmann, D., and Selbig, J. (2003b). Methods for optimizing antiviral combination therapies. *Bioinformatics*, 19 Suppl 1:i16–25.
- Beerenwinkel, N., Nolden, T., Kupfer, B., Selbig, J., Däumer, M., Hoffmann, D., Rockstroh, J., and Kaiser, R. (2001). Predicting hiv-1 cytopathogenicity and co-receptor usage from v3 envelope sequences by machine learning. In *Proceedings of the 1st IAS Conference on HIV Pathogenesis and Treatment*. Abstract 268.
- Beerenwinkel, N., Pachter, L., and Sturmfels, B. (2006). Epistasis and shapes of fitness landscapes.
- Beerenwinkel, N., Rahnenführer, J., Däumer, M., Hoffmann, D., Kaiser, R., Selbig, J., and Lengauer, T. (2005b). Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol*, 12(6):584–598.

- Beerenwinkel, N., Schmidt, B., Walter, H., Kaiser, R., Lengauer, T., Hoffmann, D., Korn, K., and Selbig, J. (2002). Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proc Natl Acad Sci U S A*, 99(12):8271–6.
- Beerenwinkel, N., Sing, T., Lengauer, T., Rahnenführer, J., Roomp, K., Savenkov, I., Fischer, R., Hoffmann, D., Selbig, J., Korn, K., Walter, H., Berg, T., Braun, P., Fätkenheuer, G., Oette, M., Rockstroh, J., Kupfer, B., Kaiser, R., and Däumer, M. (2005c). Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics*, 21(21):3943–3950.
- Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society (Series B)*, 57(1):289–300.
- Benson, C. A. (2006). Structured treatment interruptions—new findings. *Top HIV Med*, 14(3):107–111.
- Berger, E., Murphy, P., and Farber, J. (1999). Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annu Rev Immunol*, 17:657–700.
- Birkner, M., Sinisi, S., and van der Laan, M. (2005). Multiple testing and data adaptive regression: an application to HIV-1 sequence data. *Stat Appl Genet Mol Biol*, 4:Article 8.
- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Blackburn, S. (2005). *Truth: A Guide*. Oxford University Press.
- Boffito, M., Acosta, E., Burger, D., Fletcher, C. V., Flexner, C., Garaffo, R., Gatti, G., Kurowski, M., Perno, C. F., Peytavin, G., Regazzi, M., and Back, D. (2005). Therapeutic drug monitoring and drug-drug interactions involving antiretroviral drugs. *Antivir Ther*, 10(4):469–477.
- Böhm, H.-J., Klebe, G., and Kubinyi, H. (1996). *Wirkstoffdesign*. Spektrum.
- Bonhoeffer, S., Chappey, C., Parkin, N. T., Whitcomb, J. M., and Petropoulos, C. J. (2004). Evidence for positive epistasis in HIV-1. *Science*, 306(5701):1547–1550.
- Bonnet, F., Lewden, C., May, T., Heripret, L., Jouglu, E., Bevilacqua, S., Costagliola, D., Salmon, D., Chêne, G., and Morlat, P. (2004). Malignancy-related causes of death in human immunodeficiency virus-infected patients in the era of highly active antiretroviral therapy. *Cancer*, 101(2):317–324.

- Bonnet, F., Lewden, C., May, T., Heripret, L., Jouglu, E., Bevilacqua, S., Costagliola, D., Salmon, D., Chêne, G., Morlat, P., and Group, M. . S. (2005). Opportunistic infections as causes of death in HIV-infected patients in the HAART era in France. *Scand J Infect Dis*, 37(6-7):482–487.
- Branden, C. and Tooze, J. (1999). *Introduction to Protein Structure*. Routledge, 2nd edition.
- Breiman, L. (2001a). Random forests. *Machine Learning*, 45(1):5–32.
- Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215.
- Brenchley, J. M., Price, D. A., and Douek, D. C. (2006a). HIV disease: fallout from a mucosal catastrophe? *Nat Immunol*, 7(3):235–239.
- Brenchley, J. M., Price, D. A., Schacker, T. W., Asher, T. E., Silvestri, G., Rao, S., Kazzaz, Z., Bornstein, E., Lambotte, O., Altmann, D., Blazar, B. R., Rodriguez, B., Teixeira-Johnson, L., Landay, A., Martin, J. N., Hecht, F. M., Picker, L. J., Lederman, M. M., Deeks, S. G., and Douek, D. C. (2006b). Microbial translocation is a cause of systemic immune activation in chronic HIV infection. *Nat Med*, 12(12):1365–1371.
- Briggs, D., Tuttle, D., Sleasman, J., and Goodenow, M. (2000). Envelope V3 amino acid sequence predicts HIV-1 phenotype (co-receptor usage and tropism for macrophages). *AIDS*, 14(18):2937–9.
- Britton, N. F. (2003). *Essential Mathematical Biology*. Springer.
- Brumme, Z. L., Goodrich, J., Mayer, H. B., Brumme, C. J., Henrick, B. M., Wynhoven, B., Asselin, J. J., Cheung, P. K., Hogg, R. S., Montaner, J. S. G., and Harrigan, P. R. (2005a). Molecular and Clinical Epidemiology of CXCR4-Using HIV-1 in a Large Population of Antiretroviral-Naive Individuals. *J Infect Dis*, 192(3):466–74.
- Brumme, Z. L., Henrick, B. M., Brumme, C. J., Hogg, R. S., Montaner, J. S. G., and Harrigan, P. R. (2005b). Short communication. Association of the CCR5delta32 mutation with clinical response and >5-year survival following initiation of first triple antiretroviral regimen. *Antivir Ther*, 10(7):849–853.
- Brun-Vézinet, F., Costagliola, D., Khaled, M. A., Calvez, V., Clavel, F., Clotet, B., Haubrich, R., Kempf, D., King, M., Kuritzkes, D., Lanier, R., Miller, M., Miller, V., Phillips, A., Pillay, D., Schapiro, J., Scott, J., Shafer, R., Zazzi, M., Zolopa, A., and DeGruttola, V. (2004). Clinically validated genotype analysis: guiding principles and statistical concerns. *Antivir Ther*, 9(4):465–478.

- Brun-Vézinet, F., Descamps, D., Ruffault, A., Masquelier, B., Calvez, V., Peytavin, G., Telles, F., Morand-Joubert, L., Meynard, J.-L., Vray, M., Costagliola, D., and group, N. A. . S. (2003). Clinically relevant interpretation of genotype for resistance to abacavir. *AIDS*, 17(12):1795–1802.
- Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.
- Butera, S. T., editor (2005). *HIV Chemotherapy*. Caister Academic Press.
- Camacho, R. (2006). The significance of subtype-related genetic variability: controversies and unanswered questions. In Geretti, A. M., editor, *Antiretroviral resistance in clinical practice*, chapter 13, pages 117–124. Mediscript, London.
- Canutescu, A. A., Shelenkov, A. A., and Dunbrack, R. L. (2003). A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*, 12(9):2001–2014.
- Carugo, O. and Pongor, S. (2002). Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison. *J Mol Biol*, 315(4):887–898.
- Carvajal-Rodriguez, A. (2007). The importance of bio-computational tools for predicting hiv drug resistance. *Recent Patents on DNA & Gene Sequences*, 1:63–68.
- Case, T. J. (2000). *An illustrated guide to theoretical ecology*. Oxford University Press.
- CDC (1992). 1993 revised classification system for hiv infection and expanded surveillance case definition for aids among adolescents and adults. *Morbidity and Mortality Weekly Report*, 41.
- Ceccherini-Silberstein, F., Erba, F., Gago, F., Bertoli, A., Forbici, F., Bellocchi, M. C., Gori, C., D’Arrigo, R., Marcon, L., Balotta, C., Antinori, A., Monforte, A. D., and Perno, C.-F. (2004). Identification of the minimal conserved structure of HIV-1 protease in the presence and absence of drug pressure. *AIDS*, 18(12):F11–F19.
- Cejudo, J. M., Moreno, R., and Carrillo, A. (2002). Physical and neural network models of a silica-gel desiccant wheel. *Energy and Buildings*, 34(8):837–844.
- Chambers, J. M. (1993). Greater or lesser statistics: a choice for future research. *Statist. Comput.*, 3:182–184.
- Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *J Roy Stat Soc A*, 158(3):419–466.
- Check, E. (2007). Gut warfare. *Nat Med*, 13(2):116–117.

- Chesebro, B., Nishio, J., Perryman, S., Cann, A., O'Brien, W., Chen, I., and Wehrly, K. (1991). Identification of human immunodeficiency virus envelope gene sequences influencing viral entry into CD4-positive HeLa cells, T-leukemia cells, and macrophages. *J Virol*, 65(11):5782–9.
- Choe, H., Farzan, M., Sun, Y., Sullivan, N., Rollins, B., Ponath, P., Wu, L., Mackay, C., LaRosa, G., Newman, W., Gerard, N., Gerard, C., and Sodroski, J. (1996). The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. *Cell*, 85(7):1135–48.
- Choi, I.-G., Kwon, J., and Kim, S.-H. (2004). Local feature frequency profile: a method to measure structural similarity in proteins. *Proc Natl Acad Sci U S A*, 101(11):3797–3802.
- Christiansen, F. B., Otto, S. P., Bergman, A., and Feldman, M. W. (1998). Waiting with and without recombination: the time to production of a double mutant. *Theor Popul Biol*, 53(3):199–215.
- Cilliers, T., Nhlapo, J., Coetzer, M., Orlovic, D., Ketas, T., Olson, W. C., Moore, J. P., Trkola, A., and Morris, L. (2003). The CCR5 and CXCR4 coreceptors are both used by human immunodeficiency virus type 1 primary isolates from subtype C. *J Virol*, 77(7):4449–4456.
- Clavel, F. and Hance, A. J. (2004). HIV drug resistance. *N Engl J Med*, 350(10):1023–35.
- Clavel, F., Soriano, V., and Zolopa, A. (2004). *HIV Infection / Antiretroviral Resistance. Scientific basis and recommendations for management*. Bash Medical Publishing, Paris, France.
- Cocchi, F., DeVico, A., Garzino-Demo, A., Arya, S., Gallo, R., and Lusso, P. (1995). Identification of RANTES, MIP-1 alpha, and MIP-1 beta as the major HIV-suppressive factors produced by CD8+ T cells. *Science*, 270(5243):1811–5.
- Coffin, J. M. (1995). HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. *Science*, 267(5197):483–489.
- Coffin, J. M., Hughes, S. H., and Varmus, H. E. (1997). *Retroviruses*. Cold Spring Harbour Laboratory Press.
- Cohen, B. and Whitman, A. (1999). *Isaac Newton: Philosophiae Naturalis Principia Mathematica (Translation)*. University of California Press.
- Cohen, J. (2005). Is an effective HIV vaccine feasible? *Science*, 309(5731):99.

- Cohen, J. E. (2004). Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol*, 2(12):e439.
- Condra, J. H., Holder, D. J., Schleif, W. A., Blahy, O. M., Danovich, R. M., Gabryelski, L. J., Graham, D. J., Laird, D., Quintero, J. C., Rhodes, A., Robbins, H. L., Roth, E., Shivaprakash, M., Yang, T., Chodakewitz, J. A., Deutsch, P. J., Leavitt, R. Y., Massari, F. E., Mellors, J. W., Squires, K. E., Steigbigel, R. T., Tepler, H., and Emini, E. A. (1996). Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. *J Virol*, 70(12):8270–8276.
- Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*, 11(20):2463–2468.
- Costagliola, D., Cozzi-Lepri, A., Dalban, C., and Cheng, B. (2005). Initiatives for developing and comparing genotype interpretation systems step 1: external validation of existing rules-based algorithms for abacavir and ddi evaluated on virological response. *Antivir Ther*, 10:S11.
- Cozzi-Lepri, A., Ruiz, L., Loveday, C., Phillips, A. N., Clotet, B., Reiss, P., Ledergerber, B., Holkmann, C., Staszewski, S., Lundgren, J. D., and Group, E. I. D. A. S. (2005). Thymidine analogue mutation profiles: factors associated with acquiring specific profiles and their impact on the virological response to therapy. *Antivir Ther*, 10(7):791–802.
- Crandall, K. A., editor (1999). *The evolution of HIV*. Johns Hopkins University Press.
- Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.
- Cunningham, C. K., Chaix, M.-L., Rekacewicz, C., Britto, P., Rouzioux, C., Gelber, R. D., Dorenbaum, A., Delfraissy, J. F., Bazin, B., Mofenson, L., and Sullivan, J. L. (2002). Development of resistance mutations in women receiving standard antiretroviral therapy who received intrapartum nevirapine to prevent perinatal human immunodeficiency virus type 1 transmission: a substudy of pediatric AIDS clinical trials group protocol 316. *J Infect Dis*, 186(2):181–188.
- Dalglish, A., Beverley, P., Clapham, P., Crawford, D., Greaves, M., and Weiss, R. (1984). The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature*, 312(5996):763–7.
- Daly, A. K. (2007). Individualized drug therapy. *Curr Opin Drug Discov Devel*, 10(1):29–36.

- De Jong, J., De Ronde, A., Keulen, W., Tersmette, M., and Goudsmit, J. (1992). Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. *J Virol*, 66(11):6777–80.
- de Jong, J., Goudsmit, J., Keulen, W., Klaver, B., Krone, W., Tersmette, M., and de Ronde, A. (1992). Human immunodeficiency virus type 1 clones chimeric for the envelope V3 domain differ in syncytium formation and replication capacity. *J Virol*, 66(2):757–65.
- De Luca, A. (2006). The impact of resistance on viral fitness and its clinical implications. In Geretti, A. M., editor, *Antiretroviral resistance in clinical practice*, chapter 12, pages 101–116. Mediscript, London.
- De Luca, A., Cingolani, A., Giambenedetto, S. D., Trotta, M. P., Baldini, F., Rizzo, M. G., Bertoli, A., Liuzzi, G., Narciso, P., Murri, R., Ammassari, A., Perno, C. F., and Antinori, A. (2003). Variable prediction of antiretroviral treatment outcome by different systems for interpreting genotypic human immunodeficiency virus type 1 drug resistance. *J Infect Dis*, 187(12):1934–43.
- De Luca, A., Cozzi-Lepri, A., Perno, C.-F., Balotta, C., Giambenedetto, S. D., Poggio, A., Pagano, G., Tositti, G., Piscopo, R., Forno, A. D., Chiodo, F., Magnani, G., d'Arminio Monforte, A., drug resistance study group d'Arminio Monforte, I., and study group d'Arminio Monforte, I. (2004). Variability in the interpretation of transmitted genotypic HIV-1 drug resistance and prediction of virological outcomes of the initial HAART by distinct systems. *Antivir Ther*, 9(5):743–52.
- Deforche, K., Camacho, R., Grossman, Z., Silander, T., Soares, M. A., Moreau, Y., Shafer, R. W., Laethem, K. V., Carvalho, A. P., Wynhoven, B., Cane, P., Snoeck, J., Clarke, J., Sirivichayakul, S., Ariyoshi, K., Holguin, A., Rudich, H., Rodrigues, R., Bouzas, M. B., Cahn, P., Brigido, L. F., Soriano, V., Sugiura, W., Phanuphak, P., Morris, L., Weber, J., Pillay, D., Tanuri, A., Harrigan, P. R., Shapiro, J. M., Katzenstein, D. A., Kantor, R., and Vandamme, A.-M. (2007). Bayesian network analysis of resistance pathways against hiv-1 protease inhibitors. *Infect Genet Evol*, 7(3):382–390.
- Deforche, K., Silander, T., Camacho, R., Grossman, Z., Soares, M. A., Laethem, K. V., Kantor, R., Moreau, Y., Vandamme, A.-M., and non B. Workgroup (2006). Analysis of hiv-1 pol sequences using bayesian networks: implications for drug resistance. *Bioinformatics*, 22(24):2975–2979.
- DeGruttola, V., Dix, L., D'Aquila, R., Holder, D., Phillips, A., Ait-Khaled, M., Baxter, J., Clevenbergh, P., Hammer, S., Harrigan, R., Katzenstein, D., Lanier, R., Miller,

- M., Para, M., Yerly, S., Zolopa, A., Murray, J., Patick, A., Miller, V., Castillo, S., Pedneault, L., and Mellors, J. (2000). The relation between baseline HIV drug resistance and response to antiretroviral therapy: re-analysis of retrospective and prospective studies using a standardized data analysis plan. *Antivir Ther*, 5(1):41–8.
- Deng, H., Liu, R., Ellmeier, W., Choe, S., Unutmaz, D., Burkhart, M., Marzio, P. D., Marmon, S., Sutton, R. E., Hill, C. M., Davis, C. B., Peiper, S. C., Schall, T. J., Littman, D. R., and Landau, N. R. (1996). Identification of a major co-receptor for primary isolates of HIV-1. *Nature*, 381(6584):661–666.
- Derdeyn, C. A. and Hunter, E. (2005). Entry inhibitors and beyond. In *HIV Chemotherapy*. Caister Academic Press.
- Descamps, D. and Brun-Vezinet, F. (2006). *Antiretroviral resistance in clinical practice*, chapter Benefits of resistance testing, pages 73–78. Mediscript, London.
- Devereux, H. L., Emery, V. C., Johnson, M. A., and Loveday, C. (2001). Replicative fitness in vivo of hiv-1 variants with multiple drug resistance-associated mutations. *J Med Virol*, 65(2):218–224.
- DHHS Panel on Antiretroviral Guidelines for Adults and Adolescents (2006). Guidelines for the use of antiretroviral agents in hiv-1-infected adults and adolescents. Technical report, U.S. Department of Health and Human Services.
- DiRienzo, A., DeGruttola, V., Larder, B., and Hertogs, K. (2003). Non-parametric methods to predict hiv drug susceptibility phenotype from genotype. *Stat. Med.*, 22:2785–2798.
- Domingo, E., editor (2005). *Quasispecies: Concepts and Implications for Virology*. Springer.
- Domingo, E., Biebricher, C. K., Eigen, M., and Holland, J., editors (2001). *Quasispecies and RNA virus evolution: Principles and consequences*. Eureka.com / Landes Bioscience.
- Domingo, E., J.J., H., Biebricher, C., and M., E. (1995). *Quasispecies: The concept and the word.*, pages 171–180. Cambridge UP.
- Domingues, F. S., Koppensteiner, W. A., Jaritz, M., Prlic, A., Weichenberger, C., Wiederstein, M., Floeckner, H., Lackner, P., and Sippl, M. J. (1999). Sustained performance of knowledge-based potentials in fold recognition. *Proteins*, Suppl 3:112–120.

- Donaldson, Y., Bell, J., Holmes, E., Hughes, E., Brown, H., and Simmonds, P. (1994). In vivo distribution and cytopathology of variants of human immunodeficiency virus type 1 showing restricted sequence variability in the V3 loop. *J Virol*, 68(9):5991–6005.
- Doranz, B., Rucker, J., Yi, Y., Smyth, R., Samson, M., Peiper, S., Parmentier, M., Collman, R., and Doms, R. (1996). A dual-tropic primary HIV-1 isolate that uses fusin and the beta-chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. *Cell*, 85(7):1149–58.
- Dorling, D., Shaw, M., and Smith, G. D. (2006). Global inequality of life expectancy due to AIDS. *BMJ*, 332(7542):662–664.
- Dorr, P., Westby, M., Dobbs, S., Griffin, P., Irvine, B., Macartney, M., Mori, J., Rickett, G., Smith-Burchnell, C., Napier, C., Webster, R., Armour, D., Price, D., Stammen, B., Wood, A., and Perros, M. (2005). Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrob Agents Chemother*, 49(11):4721–4732.
- Douek, D. C., Picker, L. J., and Koup, R. A. (2003). T cell dynamics in HIV-1 infection. *Annu Rev Immunol*, 21:265–304.
- Draghici, S. and Potter, R. B. (2003). Predicting HIV drug resistance with neural networks. *Bioinformatics*, 19(1):98–107.
- Dragic, T., Litwin, V., Allaway, G. P., Martin, S. R., Huang, Y., Nagashima, K. A., Cayanan, C., Maddon, P. J., Koup, R. A., Moore, J. P., and Paxton, W. A. (1996). HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature*, 381(6584):667–673.
- Duhem, P. (1969). *To save the phenomena*. Chicago University Press.
- Edelstein-Keshet, L. (2004). *Mathematical Models in Biology*, volume 46 of *SIAM Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics.
- Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5):1792–1797.
- Eron, J. J., Benoit, S. L., Jemsek, J., MacArthur, R. D., Santana, J., Quinn, J. B., Kuritzkes, D. R., Fallon, M. A., and Rubin, M. (1995). Treatment with lamivudine, zidovudine, or both in hiv-positive patients with 200 to 500 cd4+ cells per cubic millimeter. north american hiv working party. *N Engl J Med*, 333(25):1662–1669.
- Faraway, J. J. (2006). *Extending the linear model with R*. Chapman & Hall / CRC.

- Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874.
- FDA (2004). Innovation or stagnation: Challenge and opportunity on the critical path to new medical products. Technical report, FDA.
- Feldman, M. W., Franklin, I., and Thomson, G. J. (1974). Selection in complex genetic systems. I. The symmetric equilibria of the three-locus symmetric viability model. *Genetics*, 76(1):135–162.
- Feldman, M. W., Otto, S. P., and Christiansen, F. B. (1996). Population genetic perspectives on the evolution of recombination. *Annu Rev Genet*, 30:261–295.
- Feng, Y., Broder, C., Kennedy, P., and Berger, E. (1996). HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science*, 272(5263):872–7. Comment: Cohen1996.
- Fenyo, E., Schuitemaker, H., B, A., and J, M. (1997). The history of hiv-1 biological phenotypes. In Korber, B., Hahn, B., Foley, B., Mellors, J., Leitner, T., Myers, G., McCutchan, F., and Kuiken, C., editors, *Human Retroviruses and AIDS 1997*, pages III–13–18. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM.
- Fine, A. (1986). Unnatural attitudes: Realist and instrumentalist attachments to science. *Mind*, 95(378):149–179.
- Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philos. Trans. Roy. Soc. London Ser. A*, 222:309–368.
- Fitzpatrick, R. (2006). A modern almagest: An updated version of ptolemy’s model of the solar system. Technical report, The University of Texas at Austin.
- Flint, S. J., Enquist, L. W., Krug, R. M., Racaniello, V. R., and Skalka, A. M. (2000). *Principles of Virology*. American Society for Microbiology.
- Fonseca, R. and Stewart, A. K. (2007). Targeted therapeutics for multiple myeloma: The arrival of a risk-stratified approach. *Mol Cancer Ther*, 6(3):802–810.
- Fouchier, R., Brouwer, M., Broersen, S., and Schuitemaker, H. (1995). Simple determination of human immunodeficiency virus type 1 syncytium-inducing V3 genotype by PCR. *J Clin Microbiol*, 33(4):906–11.
- Fouchier, R., Groenink, M., Kootstra, N., Tersmette, M., Huisman, H., Miedema, F., and Schuitemaker, H. (1992). Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J Virol*, 66(5):3183–7.

- Foulkes, A. and DeGruttola, V. (2003). Characterizing the progression of viral mutations over time. *JASA Applications and Case Studies*, 98(464):859–867.
- Freudenthal, G. (2003). 'instrumentalism' and 'realism' as categories in the history of astronomy: Duhem vs. popper, maimonides vs. gersonides. *Centaurus*, 45:227–248.
- Friedman, J., Hastie, T., and Tibshirani, R. (2000). Additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–374.
- Friedman, J. H. (1998). Data mining and statistics: What's the connection? *COMPUTING SCIENCE AND STATISTICS*, 29(1):3–9.
- Futuyma, D. J. (1998). *Evolutionary Biology*. Sinauer Associates, 3rd edition.
- Gearhart, C. A. (1985). Epicycles, eccentrics, and ellipses: The predictive capabilities of copernican planetary models. *Archive for History of Exact Sciences*, 32(3–4):207–222.
- Geretti, A. M., editor (2006). *Antiretroviral resistance in clinical practice*. Mediscript, London.
- Gillespie, J. H. (2004). *Population Genetics: A Concise Guide*. John Hopkins University Press.
- Gonzales, M. J., Wu, T. D., Taylor, J., Belitskaya, I., Kantor, R., Israelski, D., Chou, S., Zolopa, A. R., Fessel, W. J., and Shafer, R. W. (2003). Extended spectrum of HIV-1 reverse transcriptase mutations in patients receiving multiple nucleoside analog inhibitors. *AIDS*, 17(6):791–9.
- Good, R. J. (1999). Why are chemists 'turned off' by philosophy of science? *Foundations of Chemistry*, 1(2):185–215.
- Goudsmit, J., Debouck, C., Meloen, R., Smit, L., Bakker, M., Asher, D., Wolff, A., Gibbs, C., and Gajdusek, D. (1988). Human immunodeficiency virus type 1 neutralization epitope with conserved architecture elicits early type-specific antibodies in experimentally infected chimpanzees. *Proc Natl Acad Sci U S A*, 85(12):4478–82.
- Groenink, M., Andeweg, A. C., Fouchier, R. A., Broersen, S., van der Jagt, R. C., Schuitemaker, H., de Goede, R. E., Bosch, M. L., Huisman, H. G., and Tersmette, M. (1992). Phenotype-associated env gene variation among eight related human immunodeficiency virus type 1 clones: evidence for in vivo recombination and determinants of cytotropism outside the V3 domain. *J Virol*, 66(10):6175–6180.
- Grossman, Z., Meier-Schellersheim, M., Paul, W. E., and Picker, L. J. (2006). Pathogenesis of HIV infection: what the virus spares is as important as what it destroys. *Nat Med*, 12(3):289–295.

- Grote, M. N., Klitz, W., and Thomson, G. (1998). Constrained disequilibrium values and hitchhiking in a three-locus system. *Genetics*, 150(3):1295–1307.
- Guadalupe, M., Sankaran, S., George, M. D., Reay, E., Verhoeven, D., Shacklett, B. L., Flamm, J., Wegelin, J., Prindiville, T., and Dandekar, S. (2006). Viral suppression and immune restoration in the gastrointestinal mucosa of human immunodeficiency virus type 1-infected patients initiating therapy during primary or chronic infection. *J Virol*, 80(16):8236–8247.
- Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422.
- Haefner, J. (2005). *Modeling Biological Systems*. Springer.
- Hahn, B. H., Shaw, G. M., Cock, K. M. D., and Sharp, P. M. (2000). AIDS as a zoonosis: scientific and public health implications. *Science*, 287(5453):607–614.
- Hanson, N. R. (1960). The mathematical power of epicyclical astronomy. *Isis*, 51(2):150–158.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The elements of statistical learning*. Springer.
- Healy, M. J. R. (1978). Is statistics a science? *J Royal Statist. Soc. Ser. A*, 141:385–393.
- Hirsch, M. S. (1990). Chemotherapy of human immunodeficiency virus infections: current practice and future prospects. *J Infect Dis*, 161(5):845–857.
- Hoffmann, C., Rockstroh, J., and Kamps, B. S. (2006). *HIV Medicine 2006*. Flying Publisher, 14th edition.
- Hogg, R. S., Heath, K. V., Yip, B., Craib, K. J., O’Shaughnessy, M. V., Schechter, M. T., and Montaner, J. S. (1998). Improved survival among HIV-infected individuals following initiation of antiretroviral therapy. *JAMA*, 279(6):450–454.
- Hosking, J. R. M., Pednault, E. P. D., and Sudan, M. (1997). A statistical perspective on data mining. *Future Gener. Comput. Syst.*, 13(2-3):117–134.
- Huang, C., Tang, M., Zhang, M.-Y., Majeed, S., Montabana, E., Stanfield, R. L., Dimitrov, D. S., Korber, B., Sodroski, J., Wilson, I. A., Wyatt, R., and Kwong, P. D. (2005). Structure of a V3-containing HIV-1 gp120 core. *Science*, 310(5750):1025–1028.
- Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, pages 193–218.

- Hwang, S., Boyle, T., Lyerly, H., and Cullen, B. (1991). Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science*, 253(5015):71–4.
- Ismael, J. (2004). Quantum mechanics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.
- Jaakkola, T. and Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In Kearns, M. J., Solla, S. A., and Cohn, D. A., editors, *Advances in Neural Information Processing Systems 11*, pages 487–493. MIT Press, Cambridge, MA.
- Jackson, J. B., Becker-Pergola, G., Guay, L. A., Musoke, P., Mracna, M., Fowler, M. G., Mofenson, L. M., Mirochnick, M., Mmimo, F., and Eshleman, S. H. (2000). Identification of the K103N resistance mutation in Ugandan women receiving nevirapine to prevent HIV-1 vertical transmission. *AIDS*, 14(11):F111–F115.
- Jain, K. K. (2006). Challenges of drug discovery for personalized medicine. *Curr Opin Mol Ther*, 8(6):487–492.
- Jebara, T., Kondor, R., and Howard, A. (2004). Probability product kernels. *Journal of Machine Learning Research*, 5:819–844.
- Jensen, M. A., Coetzer, M., van 't Wout, A. B., Morris, L., and Mullins, J. I. (2006). A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences. *J Virol*, 80(10):4698–4704.
- Jensen, M. A., Li, F.-S., Wout, A. B. v. t., Nickle, D. C., Shriner, D., He, H.-X., McLaughlin, S., Shankarappa, R., Margolick, J. B., and Mullins, J. I. (2003). Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J Virol*, 77(24):13376–88.
- Jensen, M. A. and Wout, A. B. v. t. (2003). Predicting HIV-1 coreceptor usage with sequence analysis. *AIDS Rev*, 5(2):104–12.
- Jenwitheesuk, E. and Samudrala, R. (2005). Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antivir Ther*, 10(1):157–166.
- Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D., and Dougherty, J. P. (2000). High rate of recombination throughout the human immunodeficiency virus type 1 genome. *J Virol*, 74(3):1234–1240.

- Johnson, V. A., Brun-Vezinet, F., Clotet, B., Conway, B., Kuritzkes, D. R., Pillay, D., Schapiro, J., Telenti, A., and Richman, D. (2005). Update of the Drug Resistance Mutations in HIV-1: 2005. *Top HIV Med*, 13(1):51–7.
- Johnson, V. A., Brun-Vezinet, F., Clotet, B., Kuritzkes, D. R., Pillay, D., Schapiro, J. M., and Richman, D. D. (2006). Update of the drug resistance mutations in HIV-1: Fall 2006. *Top HIV Med*, 14(3):125–130.
- Johnston, E. R., Zijenah, L. S., Mutetwa, S., Kantor, R., Kittinunvorakoon, C., and Katzenstein, D. A. (2003). High frequency of syncytium-inducing and CXCR4-tropic viruses among human immunodeficiency virus type 1 subtype C-infected patients receiving antiretroviral treatment. *J Virol*, 77(13):7682–7688.
- Jordan, M. I. (1994). A statistical approach to decision tree modeling. In *Proc 7th Annual ACM Conference on Computational Learning Theory*, pages 13–20, New Brunswick, NJ, USA. ACM.
- Keele, B. F., Heuverswyn, F. V., Li, Y., Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Ngole, E. M., Bienvenue, Y., Delaporte, E., Brookfield, J. F. Y., Sharp, P. M., Shaw, G. M., Peeters, M., and Hahn, B. H. (2006). Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*, 313(5786):523–526.
- Kennedy, D. and Norman, C. (2005). What don't we know? *Science*, 309(5731):75.
- Klatzmann, D., Champagne, E., Chamaret, S., Gruest, J., Guetard, D., Hercend, T., Gluckman, J., and Montagnier, L. (1984). T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. *Nature*, 312(5996):767–8.
- Kleppner, D. and Jackiw, R. (2000). One hundred years of quantum physics. *Science*, 289(5481):893–898.
- Knipe, D. M., Howley, P. M., Griffin, D. E., Lamb, R. A., Martin, M. A., Roizman, B., and Straus, S. E. (2007). *Fields Virology*. Lippincott Williams & Wilkins, 5th edition.
- Koestler, A. (1989). *The Sleepwalkers: A History of Man's Changing Vision of the Universe*. Arkana, London.
- Korber, B., Farber, R., Wolpert, D., and Lapedes, A. (1993). Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc Natl Acad Sci U S A*, 90(15):7176–80.

- Korber, B., Foley, B. T., Kuiken, C., Pillai, S., and Sodroski, J. G. (1998). Numbering positions in HIV relative to HXB2. In Korber, B., editor, *The Human Retroviruses and AIDS 1998 Compendium*, pages 102–111. Los Alamos National Laboratory, Los Alamos, NM.
- Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., Hahn, B. H., Wolinsky, S., and Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288(5472):1789–1796.
- Kouyos, R. D., Otto, S. P., and Bonhoeffer, S. (2006). Effect of varying epistasis on the evolution of recombination. *Genetics*, 173(2):589–597.
- Kuhn, T. (1957). *The Copernican Revolution: Planetary Astronomy in the Development of Western Thought*. Harvard University Press.
- Kuiken, C., de Jong, J., Baan, E., Keulen, W., Tersmette, M., and Goudsmit, J. (1992). Evolution of the V3 envelope domain in proviral sequences and isolates of human immunodeficiency virus type 1 during transition of the viral biological phenotype. *J Virol*, 66(7):4622–7.
- Lafferty, J. and Lebanon, G. (2003). Information diffusion kernels. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 375–382. MIT Press, Cambridge, MA.
- Larder, B., Wang, D., and Revell, A. (2004). Previous drug exposure data significantly increase the accuracy of artificial neural networks in predicting virological response to combination therapy. *Antivir Ther*, 9:S123.
- Larder, B., Wang, D., and Revell, A. (2005). Treatment history and adherence information significantly improves prediction of virological response by neural networks. *Antivir Ther*, 10:S125.
- Larder, B., Wang, D., Revell, A., Montaner, J., Harrigan, R., De Wolf, F., Lange, J., Wegner, S., Ruiz, L., Perez-Elias, M. J., Emery, S., Gatell, J., D’Arminio Monforte, A., Torti, C., Zazzi, M., and Lane, C. (2007). The development of artificial neural networks to predict virological response to combination hiv therapy. *Antiviral Therapy*, 12(1):15–24.
- Larder, B. A., Darby, G., and Richman, D. D. (1989). HIV with reduced sensitivity to zidovudine (AZT) isolated during prolonged therapy. *Science*, 243(4899):1731–1734.
- Larder, B. A. and Kemp, S. D. (1989). Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT). *Science*, 246(4934):1155–1158.

- Lehmann, E. L. (1990). Model specification: The views of Fisher and Neyman and later developments. *Statistical Science*, 5:160–168.
- Lengauer, T. and Sing, T. (2006). Bioinformatics-assisted anti-HIV therapy. *Nat Rev Microbiol*, 4(10):790–797.
- Lenski, R. E., Winkworth, C. L., and Riley, M. A. (2003). Rates of DNA sequence evolution in experimental populations of *Escherichia coli* during 20,000 generations. *J Mol Evol*, 56(4):498–508.
- Logie, D. (1999). AIDS cuts life expectancy insub-saharan Africa by a quarter. *BMJ*, 319(7213):806.
- Lu, L., Lu, H., and Skolnick, J. (2002). MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. *Proteins*, 49(3):350–364.
- Lucas, P. (2004). Bayesian analysis, pattern analysis, and data mining in health care. *Curr Opin Crit Care*, 10(5):399–403.
- Lusso, P. (2006). HIV and the chemokine system: 10 years later. *EMBO J*, 25(3):447–456.
- Mansky, L. M. and Temin, H. M. (1995). Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. *J Virol*, 69(8):5087–5094.
- Marcelin, A.-G. (2006). Resistance to nucleoside reverse transcriptase inhibitors. In Geretti, A. M., editor, *Antiretroviral resistance in clinical practice*, pages 1–10. Mediscript.
- Markel, H. (2005). The search for effective hiv vaccines. *N Engl J Med*, 353(8):753–757.
- Mathers, C. D. and Loncar, D. (2006). Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med*, 3(11):e442.
- Matthews, T., Salgo, M., Greenberg, M., Chung, J., DeMasi, R., and Bolognesi, D. (2004). Enfuvirtide: the first therapy to inhibit the entry of HIV-1 into host CD4 lymphocytes. *Nat Rev Drug Discov*, 3(3):215–225.
- Mehandru, S., Poles, M. A., Tenner-Racz, K., Jean-Pierre, P., Manuelli, V., Lopez, P., Shet, A., Low, A., Mohri, H., Boden, D., Racz, P., and Markowitz, M. (2006). Lack of Mucosal Immune Reconstitution during Prolonged Treatment of Acute and Early HIV-1 Infection. *PLoS Med*, 3(12):e484.

- Meyer, D., Leisch, F., and Hornik, K. (2002). Benchmarking support vector machines. Technical Report 78, Sonderforschungsbereich "Adaptive Information Systems and Modelling in Economics and Management Science".
- Milich, L., Margolin, B., and Swanstrom, R. (1997). Patterns of amino acid variability in NSI-like and SI-like V3 sequences and a linked change in the CD4-binding domain of the HIV-1 Env protein. *Virology*, 239(1):108–18.
- Mintseris, J. and Weng, Z. (2003). Atomic contact vectors in protein-protein recognition. *Proteins*, 53(3):629–639.
- Mitchell, T. (1997). *Machine Learning*. McGraw Hill.
- Mocroft, A., Vella, S., Benfield, T. L., Chiesi, A., Miller, V., Gargalianos, P., d'Arminio Monforte, A., Yust, I., Bruun, J. N., Phillips, A. N., and Lundgren, J. D. (1998). Changing patterns of mortality across Europe in patients infected with HIV-1. EuroSIDA Study Group. *Lancet*, 352(9142):1725–1730.
- Modrow, S., Hahn, B., Shaw, G., Gallo, R., Wong-Staal, F., and Wolf, H. (1987). Computer-assisted analysis of envelope protein sequences of seven human immunodeficiency virus isolates: prediction of antigenic epitopes in conserved and variable regions. *J Virol*, 61(2):570–8.
- Moore, J. P., Kitchen, S. G., Pugach, P., and Zack, J. A. (2004). The CCR5 and CXCR4 coreceptors—central to understanding the transmission and pathogenesis of human immunodeficiency virus type 1 infection. *AIDS Res Hum Retroviruses*, 20(1):111–126.
- Murray, J. (2002). *Mathematical Biology. I: An Introduction*. Springer.
- Murray, J. (2004). *Mathematical Biology. II: Spatial Models and Biomedical Applications*. Springer.
- Neyman, J. (1939). On a new class of 'contagious' distributions, applicable in entomology and bacteriology. *Ann. Math. Statist.*, 10:35–57.
- Nicholson, J. K. (2006). Global systems biology, personalized medicine and molecular epidemiology. *Mol Syst Biol*, 2:52.
- Niiniluoto, I. (2007). Scientific progress. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*.
- Nosten, F. and Brasseur, P. (2002). Combination therapy for malaria: the way forward? *Drugs*, 62(9):1315–1329.

- Nowak, M. A. (2006). *Evolutionary Dynamics*. Harvard University Press.
- Nowak, M. A. and May, R. (2000). *Virus dynamics: Mathematical principles of immunology and virology*. Oxford University Press.
- O'Connor, J. J. and Robertson, E. F. (2003). Mathematics and the physical world. In *The MacTutor History of Mathematics Archive*. University of St. Andrews, Scotland. <http://www-groups.dcs.st-and.ac.uk/~history/HistTopics/World.html>.
- Opar, A. (2007). New HIV drug classes on the horizon. *Nature Reviews Drug Discovery*, 6(4):258–259.
- Otto, S. P. and Lenormand, T. (2002). Resolving the paradox of sex and recombination. *Nat Rev Genet*, 3(4):252–261.
- Ozols, R. F., Herbst, R. S., Colson, Y. L., Gralow, J., Bonner, J., Curran, W. J., Eisenberg, B. L., Ganz, P. A., Kramer, B. S., Kris, M. G., Markman, M., Mayer, R. J., Raghavan, D., Reaman, G. H., Sawaya, R., Schilsky, R. L., Schuchter, L. M., Sweetenham, J. W., Vahdat, L. T., Winn, R. J., and of Clinical Oncology, A. S. (2007). Clinical cancer advances 2006: major research advances in cancer treatment, prevention, and screening—a report from the American Society of Clinical Oncology. *J Clin Oncol*, 25(1):146–162.
- Palella, F. J., Delaney, K. M., Moorman, A. C., Loveless, M. O., Fuhrer, J., Satten, G. A., Aschman, D. J., and Holmberg, S. D. (1998). Declining morbidity and mortality among patients with advanced human immunodeficiency virus infection. HIV Outpatient Study Investigators. *N Engl J Med*, 338(13):853–860.
- Palumbi, S. R. (2001). Humans as the world's greatest evolutionary force. *Science*, 293(5536):1786–1790.
- Papadopoulos, D., Schneider, D., Meier-Eiss, J., Arber, W., Lenski, R. E., and Blot, M. (1999). Genomic evolution during a 10,000-generation experiment with bacteria. *Proc Natl Acad Sci U S A*, 96(7):3807–3812.
- Pastore, C., Nedellec, R., Ramos, A., Pontow, S., Ratner, L., and Mosier, D. E. (2006). Human immunodeficiency virus type 1 coreceptor switching: V1/V2 gain-of-fitness mutations compensate for V3 loss-of-fitness mutations. *J Virol*, 80(2):750–758.
- Paterlini, M. G. (2002). Structure modeling of the chemokine receptor CCR5: implications for ligand binding and selectivity. *Biophys J*, 83(6):3012–31.

- Penn, M. L., Grivel, J. C., Schramm, B., Goldsmith, M. A., and Margolis, L. (1999). CXCR4 utilization is sufficient to trigger CD4+ T cell depletion in HIV-1-infected human lymphoid tissue. *Proc Natl Acad Sci U S A*, 96(2):663–668.
- Perno, C. F., Cozzi-Lepri, A., Balotta, C., Forbici, F., Violin, M., Bertoli, A., Facchi, G., Pezzotti, P., Cadeo, G., Tositti, G., Pasquinucci, S., Pauluzzi, S., Scalzini, A., Salassa, B., Vincenti, A., Phillips, A. N., Dianzani, F., Appice, A., Angarano, G., Monno, L., Ippolito, G., Moroni, M., d' Arminio Monforte, A., and Group, I. C. N. A. I. S. (2001). Secondary mutations in the protease region of human immunodeficiency virus and virologic failure in drug-naive patients treated with protease inhibitor-based therapy. *J Infect Dis*, 184(8):983–991.
- Pillai, S., Good, B., Richman, D., and Corbeil, J. (2003). A new perspective on V3 phenotype prediction. *AIDS Res Hum Retroviruses*, 19(2):145–9.
- Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. Springer.
- Platt, J. (1999). *Advances in Large Margin Classifiers*, chapter Probabilities for Support Vector Machines, pages 61–74. MIT Press.
- Pontil, M., Mukherjee, S., and Girosi, F. (2000). On the noise model of support vector machines regression. In *ALT '00: Proceedings of the 11th International Conference on Algorithmic Learning Theory*, pages 316–324, London, UK. Springer-Verlag.
- Popovic, M., Sarngadharan, M. G., Read, E., and Gallo, R. C. (1984). Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science*, 224(4648):497–500.
- Popper, K. (1959). *The logic of scientific discovery*. Basic Books, New York.
- Popper, K. (1968). Three views concerning human knowledge. In *Conjectures and refutations. The growth of scientific knowledge*. Harper and Row, New York.
- Poveda, E. and Soriano, V. (2006). Resistance to entry inhibitors. In Geretti, A. M., editor, *Antiretroviral resistance in clinical practice*, pages 25–32. Mediscript.
- Prosperi, M., Di Giambenedetto, S., and Trotta, M. (2004). A fuzzy relational system trained by genetic algorithms and HIV-1 resistance genotypes/virological response data from prospective studies usefully predicts treatment outcomes. *Antivir Ther*, 9:S121.
- Prosperi, M., Zazzi, M., and Perno, C. (2005). 'Common law' applied to treatment decisions for drug resistant HIV. *Antivir Ther*, 10:S62.

- Rabinowitz, M., Myers, L., Banjevic, M., Chan, A., Sweetkind-Singer, J., Haberer, J., McCann, K., and Wolkowicz, R. (2006). Accurate prediction of HIV-1 drug response from the reverse transcriptase and protease amino acid sequences using sparse models created by convex optimization. *Bioinformatics*, 22(5):541–549.
- Rahnenfuhrer, J., Beerenwinkel, N., Schulz, W. A., Hartmann, C., von Deimling, A., Wullich, B., and Lengauer, T. (2005). Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics*, 21(10):2438–46.
- Reed, H. S. and Holland, R. H. (1919). The Growth Rate of an Annual Plant *Helianthus*. *Proc Natl Acad Sci U S A*, 5(4):135–144.
- Regoes, R. R. and Bonhoeffer, S. (2005). The HIV coreceptor switch: a population dynamical perspective. *Trends Microbiol*, 13(6):269–277.
- Resch, W., Hoffman, N., and Swanstrom, R. (2001). Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology*, 288(1):51–62.
- Richman, D. D., Havlir, D., Corbeil, J., Looney, D., Ignacio, C., Spector, S. A., Sullivan, J., Cheeseman, S., Barringer, K., and Pauletti, D. (1994). Nevirapine resistance mutations of human immunodeficiency virus type 1 selected during therapy. *J Virol*, 68(3):1660–1666.
- Rodrigo, A. and Learn, G. H., editors (2001). *Computational and evolutionary analysis of HIV molecular sequences*. Kluwer Academic Publishers.
- Rodriguez, B., Sethi, A. K., Cheruvu, V. K., Mackay, W., Bosch, R. J., Kitahata, M., Boswell, S. L., Mathews, W. C., Bangsberg, D. R., Martin, J., Whalen, C. C., Sieg, S., Yadavalli, S., Deeks, S. G., and Lederman, M. M. (2006). Predictive value of plasma HIV RNA level on rate of CD4 T-cell decline in untreated HIV infection. *JAMA*, 296(12):1498–1506.
- Rosen, O., Samson, A. O., Sharon, M., Zolla-Pazner, S., and Anglistter, J. (2006). Response to matters arising. *Structure*, 14(4):649–651.
- Rosenberg, A. (2000). *The Philosophy of Science: a contemporary introduction*. Routledge.
- Ruppert, D., Wand, M. P., and Carroll, R. J. (2005). *Semiparametric regression*. Cambridge University Press.
- Ruschhaupt, M., Huber, W., Poustka, A., and Mansmann, U. (2004). A compendium to ensure computational reproducibility in high-dimensional classification tasks. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 37.

- Sabin, C. (2006). *Antiretroviral resistance in clinical practice*, chapter Database analyses of predictors of resistance, pages 125–132. Mediscript, London.
- Sagan, B. E. (2001). *The symmetric group*. Springer, 2nd edition.
- Sammon, J. (1969). A non-linear mapping for data structure analysis. *IEEE Trans. Comput.*, C-18(5):401–409.
- Sander, O., Sing, T., Sommer, I., Low, A. J., Cheung, P. K., Harrigan, P. R., Lengauer, T., and Domingues, F. S. (2007). Structural Descriptors of gp120 V3 Loop for the Prediction of HIV-1 Coreceptor Usage. *PLoS Comput Biol*, 3(3):e58.
- Schackman, B. R., Gebo, K. A., Walensky, R. P., Losina, E., Muccio, T., Sax, P. E., Weinstein, M. C., Seage, G. R., Moore, R. D., and Freedberg, K. A. (2006). The lifetime cost of current human immunodeficiency virus care in the United States. *Med Care*, 44(11):990–997.
- Scheib, H., Sperisen, P., and Hartley, O. (2006). HIV-1 coreceptor selectivity: structural analogy between HIV-1 V3 regions and chemokine beta-hairpins is not the explanation. *Structure*, 14(4):645–7; discussion 649–51.
- Schirrmacher, A. (2003). Planting in his neighbor’s garden: David Hilbert and early Göttingen quantum physics. *Phys. Perspect.*, 5:4–20.
- Schmidt, B., Walter, H., Zeitler, N., and Korn, K. (2002). Genotypic drug resistance interpretation systems—the cutting edge of antiretroviral therapy. *AIDS Rev*, 4(3):148–156.
- Schmit, J. C., Ruiz, L., Stuyver, L., Laethem, K. V., Vanderlinden, I., Puig, T., Rossau, R., Desmyter, J., Clercq, E. D., Clotet, B., and Vandamme, A. M. (1998). Comparison of the LiPA HIV-1 RT test, selective PCR and direct solid phase sequencing for the detection of HIV-1 drug resistance mutations. *J Virol Methods*, 73(1):77–82.
- Schmitt, S., Kuhn, D., and Klebe, G. (2002). A new method to detect related function among proteins independent of sequence and fold homology. *J Mol Biol*, 323(2):387–406.
- Schölkopf, B. and Smola, A. (2002). *Learning with Kernels*. MIT Press.
- Schölkopf, B., Tsuda, K., and Vert, J.-P., editors (2004). *Kernel methods in computational biology*. MIT Press, Cambridge, MA.
- Schuitemaker, H., Kootstra, N. A., de Goede, R. E., de Wolf, F., Miedema, F., and Tersmette, M. (1991). Monocytotropic human immunodeficiency virus type 1 (HIV-1) variants detectable in all stages of HIV-1 infection lack T-cell line tropism and syncytium-inducing ability in primary T-cell culture. *J Virol*, 65(1):356–363.

- Schutten, M. (2006). Resistance assays. In Geretti, A. M., editor, *Antiretroviral resistance in clinical practice*, pages 33–42. Mediscript.
- Segal, M. R., Barbour, J. D., and Grant, R. M. (2004). Relating hiv-1 sequence variation to replication capacity via trees and forests. *Statistical Applications in Genetics and Molecular Biology*, 3(1):Article 2.
- Service, R. F. (2006). Gene sequencing. The race for the \$1000 genome. *Science*, 311(5767):1544–1546.
- Sevin, A. D., DeGruttola, V., Nijhuis, M., Schapiro, J. M., Foulkes, A. S., Para, M. F., and Boucher, C. A. (2000). Methods for investigation of the relationship between drug-susceptibility phenotype and human immunodeficiency virus type 1 genotype with applications to AIDS clinical trials group 333. *J Infect Dis*, 182(1):59–67.
- Shafer, R. W. and Schapiro, J. M. (2005). Drug resistance and antiretroviral drug development. *J Antimicrob Chemother*, 55(6):817–820.
- Shankarappa, R., Margolick, J., Gange, S., Rodrigo, A., Upchurch, D., Farzadegan, H., Gupta, P., Rinaldo, C., Learn, G., He, X., Huang, X., and Mullins, J. (1999). Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. *J Virol*, 73(12):10489–502.
- Sharon, M., Kessler, N., Levy, R., Zolla-Pazner, S., Görlach, M., and Anglister, J. (2003). Alternative conformations of HIV-1 V3 loops mimic beta hairpins in chemokines, suggesting a mechanism for coreceptor selectivity. *Structure (Camb)*, 11(2):225–36.
- Sharp, P. M., Bailes, E., Chaudhuri, R. R., Rodenburg, C. M., Santiago, M. O., and Hahn, B. H. (2001). The origins of acquired immune deficiency syndrome viruses: where and when? *Philos Trans R Soc Lond B Biol Sci*, 356(1410):867–876.
- Shenderovich, M. D., Kagan, R. M., Heseltine, P. N. R., and Ramnarayan, K. (2003). Structure-based phenotyping predicts HIV-1 protease inhibitor resistance. *Protein Sci*, 12(8):1706–1718.
- Simon, V. and Ho, D. D. (2003). HIV-1 dynamics in vivo: implications for therapy. *Nat Rev Microbiol*, 1(3):181–190.
- Simon, V., Ho, D. D., and Karim, Q. A. (2006). HIV/AIDS epidemiology, pathogenesis, prevention, and treatment. *Lancet*, 368(9534):489–504.
- Sing, T. (2004). Learning localized rule mixtures by maximizing the area under the ROC curve, with an application the the prediction of HIV-1 coreceptor usage. Master's thesis, Albert-Ludwigs-Universität Freiburg (Germany).

- Sing, T. and Beerenwinkel, N. (2007). Mutagenetic tree Fisher kernel improves prediction of drug resistance from viral genotype. In *Advances in Neural Information Processing Systems 19*. In press.
- Sing, T., Beerenwinkel, N., and Lengauer, T. (2004). Learning mixtures of localized rules by maximizing the area under the ROC curve. In Hernández-Orallo, J., Ferri, C., Lachiche, N., and Flach, P. A., editors, *1st Int. Workshop on ROC Analysis in Artificial Intelligence*, pages 89–96, Valencia, Spain.
- Sing, T., Low, A., Beerenwinkel, N., Sander, O., Cheung, P., Domingues, F., Büch, J., Däumer, M., Kaiser, R., Lengauer, T., and Harrigan, P. (2007). Predicting HIV co-receptor usage based on genetic and clinical covariates. *Antivir Ther*, page (in press).
- Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005a). ROCr: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941.
- Sing, T., Svicher, V., Beerenwinkel, N., Ceccherini-Silberstein, F., Däumer, M., Kaiser, R., Walter, H., Korn, K., Hoffmann, D., Oette, M., Rockstroh, J. K., Fätkenheuer, G., Perno, C.-F., and Lengauer, T. (2005b). Characterization of novel HIV drug resistance mutations using clustering, multidimensional scaling, and SVM-based feature ranking. In Jorge, A. and Torgo, L., editors, *Knowledge Discovery in Databases: PKDD 2005*, Lecture Notes in Artificial Intelligence, pages 285–296. Springer.
- Sing, T., Tschochner, M., Weisser, H., Verheyen, J., Daumer, M., Kaiser, R., Korn, K., Walter, H., and Lengauer, T. (2006). Predictive and interpretable models relating hiv-1 gag/pol genotype to replication capacity. In *4th European HIV Drug Resistance Workshop*, Monaco, Monte Carlo. Abstract 55.
- Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr Opin Struct Biol*, 5(2):229–235.
- Sirois, S., Sing, T., and Chou, K.-C. (2005). HIV-1 gp120 V3 loop for structure-based drug design. *Curr Protein Pept Sci*, 6(5):413–422.
- Skrabal, K., Low, A. J., Dong, W., Sing, T., Cheung, P. K., Mammano, F., and Harrigan, P. R. (2007). Determining human immunodeficiency virus coreceptor use in a clinical setting: degree of correlation between two phenotypic assays and a bioinformatic model. *J Clin Microbiol*, 45(2):279–284.
- Smith, K. (2003). The HIV vaccine saga. *Med Immunol*, 2(1):1.
- Sompayrac, L. (2002). *How pathogenic viruses work*. Blackwell, Malden, MA.

- Stahl, M., Taroni, C., and Schneider, G. (2000). Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. *Protein Eng*, 13(2):83–88.
- Svicher, V., Ceccherini-Silberstein, F., T.Sing, Santoro, M., Beerenwinkel, N., Rodríguez, F., Forbici, F., d'Arminio Monforte, A., Antinori, A., and Perno, C. (2005). Additional mutations in hiv-1 reverse transcriptase are involved in the highly ordered regulation of nrti resistance. In *Proc. 3rd European HIV Drug Resistance Workshop*, Athens, Greece. Abstract 63. **Best oral presentation award.**
- Svicher, V., Sing, T., Santoro, M. M., Forbici, F., Rodríguez-Barrios, F., Bertoli, A., Beerenwinkel, N., Bellocchi, M. C., Gago, F., d'Arminio Monforte, A., Antinori, A., Lengauer, T., Ceccherini-Silberstein, F., and Perno, C. F. (2006). Involvement of novel human immunodeficiency virus type 1 reverse transcriptase mutations in the regulation of resistance to nucleoside inhibitors. *J Virol*, 80(14):7186–7198. **V.S. and T.S. contributed equally to this work.**
- Swanstrom, R., Bosch, R. J., Katzenstein, D., Cheng, H., Jiang, H., Hellmann, N., Haubrich, R., Fiscus, S. A., Fletcher, C. V., Acosta, E. P., Gulick, R. M., and Gulick, A. C. T. G. P. . T. (2004). Weighted Phenotypic Susceptibility Scores Are Predictive of the HIV-1 RNA Response in Protease Inhibitor-Experienced HIV-1-Infected Subjects. *J Infect Dis*, 190(5):886–93.
- Tan, W.-Y. and Wu, H., editors (2005). *Deterministic and Stochastic Models of AIDS Epidemics and HIV Infection with Intervention*. World Scientific Publishing.
- Taylor, R. G. (1998). *Models of Computation and Formal Languages*. Oxford University Press.
- Temin, H. M. and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*, 226(5252):1211–1213.
- Tersmette, M., Lange, J. M., de Goede, R. E., de Wolf, F., Eeftink-Schattenkerk, J. K., Schellekens, P. T., Coutinho, R. A., Huisman, J. G., Goudsmit, J., and Miedema, F. (1989). Association between biological properties of human immunodeficiency virus variants and risk for aids and aids mortality. *Lancet*, 1(8645):983–985.
- Therneau, T. M. (2001). *Modeling Survival Data: Extending the Cox Model*. Springer.
- Tsuda, K., Kawanabe, M., Rätsch, G., Sonnenburg, S., and Müller, K. (2002a). A new discriminative kernel from probabilistic models. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems 14*, pages 977–984. MIT Press, Cambridge, MA.

- Tsuda, K., Kin, T., and Asai, K. (2002b). Marginalized kernels for biological sequences. *Bioinformatics*, 18 Suppl 1:S268–S275.
- UNAIDS (2006). Aids epidemic update. Technical report, United Nations Programme on HIV/AIDS.
- United States (2005). *Economic Report of the President*. United States Government Printing Office, Washington.
- van der Greef, J., Hankemeier, T., and McBurney, R. N. (2006). Metabolomics-based systems biology and personalized medicine: moving towards n = 1 clinical trials? *Pharmacogenomics*, 7(7):1087–1094.
- Van Houtte, M., Vermeiren, H., and Lecocq, P. (2004). The virtualphenotype analysis of an hiv-1 genotype provides a more accurate prediction of drug susceptibility than a single phenotype measurement. In *13th International Symposiums on HIV and Emerging Infectious Diseases*, page OP. 8.9, Toulon.
- Van Laethem, K., Luca, A. D., Antinori, A., Cingolani, A., Perna, C. F., and Vandamme, A.-M. (2002). A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in hiv-1-infected patients. *Antivir Ther*, 7(2):123–129.
- Vapnik, V. (1998). *Statistical learning theory*.
- Vapnik, V. (2005). Universal learning technology: support vector machines. *NEC Journal of Advanced Technology*, 2(2):137–144.
- Vermeiren, H., Van den Bulcke, T., and Van Marck, H. (2004). Application of multiple linear regression modelling to the quantitative prediction of HIV-1 drug susceptibility phenotype from viral genotype. *Antivir Ther*, 9:S122.
- Walter, H., Schmidt, B., Korn, K., Vandamme, A. M., Harrer, T., and Uberla, K. (1999). Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors. *J Clin Virol*, 13(1-2):71–80.
- Wang, D. and Larder, B. (2003). Enhanced prediction of lopinavir resistance from genotype by use of artificial neural networks. *J Infect Dis*, 188(5):653–660.
- Wang, K., Jenwitheesuk, E., Samudrala, R., and Mittler, J. (2004a). Simple linear model provides highly accurate genotypic predictions of hiv-1 drug resistance. *Antivir Ther*, 9(3):343–352.
- Wang, K., Samudrala, R., and Mittler, J. E. (2004b). HIV-1 genotypic drug-resistance interpretation algorithms need to include hypersusceptibility-associated mutations. *J Infect Dis*, 190(11):2055–6.

- Wasserman, L. (2004). *All of statistics*. Springer.
- Watabe, T., Kishino, H., Okuhara, Y., and Kitazoe, Y. (2006). Fold recognition of the human immunodeficiency virus type 1 V3 loop and flexibility of its crown structure during the course of adaptation to a host. *Genetics*, 172(3):1385–1396.
- Weinreich, D. M. (2005). The rank ordering of genotypic fitness values predicts genetic constraint on natural selection on landscapes lacking sign epistasis. *Genetics*, 171(3):1397–1405.
- Weisser, H. (2006). Predictive and interpretable models relating hiv-1 pol genotype to replication capacity (bachelor thesis). Technical report, Saarland University.
- West, M. L. and Fairlie, D. P. (1995). Targeting HIV-1 protease: a test of drug-design methodologies. *Trends Pharmacol Sci*, 16(2):67–75.
- Westby, M. and van der Ryst, E. (2005). CCR5 antagonists: host-targeted antivirals for the treatment of HIV infection. *Antivir Chem Chemother*, 16(6):339–354.
- Westfall, R. S. (1978). *The Construction of Modern Science: Mechanisms and Mechanics*. Cambridge University Press.
- Wild, C. J. (1994). Embracing the "wider view" of statistics. *Am. Statistician*, 48:163–171.
- Wolpert, D. (2001). The supervised learning no-free-lunch theorems. In *Proc. 6th Online World Conference on Soft Computing in Industrial Applications*.
- World Health Organization (2005). *Basic Documents*, chapter Constitution of the World Health Organization, pages 1–18. WHO Marketing and Dissemination, 45th edition.
- Wu, T. D., Schiffer, C. A., Gonzales, M. J., Taylor, J., Kantor, R., Chou, S., Israelski, D., Zolopa, A. R., Fessel, W. J., and Shafer, R. W. (2003). Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *J Virol*, 77(8):4836–47.
- Wynne, S. M. and Davey, R. T. (2005). Immune based therapies in the treatment of hiv-1 infection. In Butera, S. T., editor, *HIV Chemotherapy: A Critical Review*. Caister Academic Press, Norwich, UK.
- Xiao, L., Owen, S., Goldman, I., Lal, A., deJong, J., Goudsmit, J., and Lal, R. (1998). CCR5 coreceptor usage of non-syncytium-inducing primary HIV-1 is independent of phylogenetically distinct global HIV-1 isolates: delineation of consensus motif in the V3 domain that predicts CCR-5 usage. *Virology*, 240(1):83–92.

- Zazzi, M. (2006). Euresist standard datum. Technical report, EuResist Consortium. Available at [www.euresist.org/contenuti/D2\\_1.pdf](http://www.euresist.org/contenuti/D2_1.pdf).
- Zhang, H., Ortí, G., Du, Q., He, J., Kankasa, C., Bhat, G., and Wood, C. (2002). Phylogenetic and phenotypic analysis of hiv type 1 env gp120 in cases of subtype c mother-to-child transmission. *AIDS Res Hum Retroviruses*, 18(18):1415–1423.
- Zhang, Y., Devries, M. E., and Skolnick, J. (2006). Structure modeling of all identified g protein-coupled receptors in the human genome. *PLoS Comput Biol*, 2(2):e13.
- Zhong, P., Burda, S., Konings, F., Urbanski, M., Ma, L., Zekeng, L., Ewane, L., Agyingi, L., Agwara, M., Saa, Afane, Z. E., Kinge, T., Zolla-Pazner, S., and Nyambi, P. (2003). Genetic and biological properties of hiv type 1 isolates prevalent in villagers of the cameroon equatorial rain forests and grass fields: further evidence of broad hiv type 1 genetic diversity. *AIDS Res Hum Retroviruses*, 19(12):1167–1178.
- Zhuang, J., Jetzt, A. E., Sun, G., Yu, H., Klarmann, G., Ron, Y., Preston, B. D., and Dougherty, J. P. (2002). Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J Virol*, 76(22):11273–11282.
- Zolopa, A., Kiazand, A., and Race, E. (2004). Resistance tests interpretation. In Clavel, F., Soriano, V., and Zolopa, A., editors, *HIV Infection: Antiretroviral Resistance*, pages 77–92. Bash Medical Publishing, Paris.

---

# Appendix: List of Publications

The work summarized in this thesis spans five conference papers, 17 journal papers, one book chapter, and more than 40 conference abstracts/posters. These publications are listed below — except for conference abstracts/posters, where the list is limited to award-winning contributions.

## Papers in conference proceedings

Roomp, K., Beerenwinkel, N., Sing, T., Schülter, E., Büch, J., Sierra-Aragon, S., Däumer, M., Hoffmann, D., Kaiser, R., Lengauer, T., and Selbig, J. (2006). Arevir: A secure platform for designing personalized antiretroviral therapies against hiv. In *Data Integration in the Life Sciences, Third International Workshop, DILS 2006*, volume 4075 of *Lecture Notes in Computer Science*, pages 185–194. Springer.

Sing, T. and Beerenwinkel, N. (2007). Mutagenetic tree Fisher kernel improves prediction of drug resistance from viral genotype. In *Advances in Neural Information Processing Systems 19*. In press.

Sing, T., Beerenwinkel, N., and Lengauer, T. (2004). Learning mixtures of localized rules by maximizing the area under the ROC curve. In Hernández-Orallo, J., Ferri, C., Lachiche, N., and Flach, P. A., editors, *1st Int. Workshop on ROC Analysis in Artificial Intelligence*, pages 89–96, Valencia, Spain.

Sing, T., Beerenwinkel, N., Lengauer, T., Svicher, V., Ceccherini-Silberstein, F., and Perno, C.-F. (2005a). Supervised and unsupervised approaches for analyzing covariation and phenotypic impact of hiv drug resistance mutations. In *LWA 2005, Lernen Wissensentdeckung Adaptivität*, pages 234–239, Saarland University, Saarbrücken, Germany. German Research Center for Artificial Intelligence (DFKI).

Sing, T., Svicher, V., Beerenwinkel, N., Ceccherini-Silberstein, F., Däumer, M., Kaiser, R., Walter, H., Korn, K., Hoffmann, D., Oette, M., Rockstroh, J. K., Fätkenheuer, G., Perno, C.-F., and Lengauer, T. (2005b). Characterization of novel HIV

drug resistance mutations using clustering, multidimensional scaling, and SVM-based feature ranking. In Jorge, A. and Torgo, L., editors, *Knowledge Discovery in Databases: PKDD 2005*, Lecture Notes in Artificial Intelligence, pages 285–296. Springer.

## Journal articles

Altmann, A., Beerenwinkel, N., Sing, T., Savenkov, I., Däumer, M., Kaiser, R., Rhee, S.-Y., Fessel, W. J., Shafer, R. W., and Lengauer, T. (2007). Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antiviral Therapy*, 12(2):169–178.

Beerenwinkel, N., Däumer, M., Sing, T., Rahnenführer, J., Lengauer, T., Selbig, J., Hoffmann, D., and Kaiser, R. (2005a). Estimating HIV evolutionary pathways and the genetic barrier to drug resistance. *J Infect Dis*, 191(11):1953–1960.

Beerenwinkel, N., Sing, T., Lengauer, T., Rahnenführer, J., Roomp, K., Savenkov, I., Fischer, R., Hoffmann, D., Selbig, J., Korn, K., Walter, H., Berg, T., Braun, P., Fätkenheuer, G., Oette, M., Rockstroh, J., Kupfer, B., Kaiser, R., and Däumer, M. (2005b). Computational methods for the design of effective therapies against drug resistant HIV strains. *Bioinformatics*, 21(21):3943–3950.

Ceccherini-Silberstein, F., Svicher, V., Sing, T., Artese, A., Santoro, M. M., Gori, C., Bertoli, A., Alcaro, S., Palamara, G., d'Arminio Monforte, A., Antinori, A., Lengauer, T., and Perno, C. F. (2007). Additional mutations in HIV-1 reverse transcriptase involved in the regulation of resistance to non-nucleoside inhibitors. Submitted.

Kupfer, B., Sing, T., Schöffler, P., Hall, R., Kurz, R., McKeown, A., Schneeweis, K.-E., Eberl, W., Oldenburg, J., Brackmann, H. H., Rockstroh, J. K., Spengler, U., Däumer, M. P., Kaiser, R., Lengauer, T., and Matz, B. (2007). Fifteen years of env C2-V3-C3 evolution in six individuals clonally infected with human immunodeficiency virus type 1. *J. Med. Virol.*, page in press. Submitted.

Lehmann, C., Däumer, M., Boussaad, I., Sing, T., Beerenwinkel, N., Lengauer, T., Schmeisser, N., Wyen, C., Fätkenheuer, G., and Kaiser, R. (2006). Stable coreceptor usage of HIV in patients with ongoing treatment failure on HAART. *J Clin Virol*, 37(4):300–304.

Lengauer, T. and Sing, T. (2006). Bioinformatics-assisted anti-HIV therapy. *Nat Rev Microbiol*, 4(10):790–797.

Low, A. J., Dong, W., Chan, D., Sing, T., Swanstrom, R., Jensen, M., Pillai, S., Good, B., and Harrigan, P. R. (2007). Current implementations of V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. *AIDS*, page in press. Submitted.

Sander, O., Sing, T., Sommer, I., Low, A. J., Cheung, P. K., Harrigan, P. R., Lengauer, T., and Domingues, F. S. (2007). Structural Descriptors of gp120 V3 Loop for the Prediction of HIV-1 Coreceptor Usage. *PLoS Comput Biol*, 3(3):e58.

Sing, T., Altmann, A., Vermeiren, H., Winters, B., Van Craenenbroeck, E., Van der Borght, K., Rhee, S., Shafer, R., Schülter, E., Kaiser, R., Peres, Y., Sonnerborg, A., Fessel, W. J., Incardona, F., Zazzi, M., Bachelier, L., Van Vlijmen, H., and Lengauer, T. (2007a). Inferring virological response from genotype: With or without predicted phenotypes? Submitted.

Sing, T., Low, A., Beerenwinkel, N., Sander, O., Cheung, P., Domingues, F., Büch, J., Däumer, M., Kaiser, R., Lengauer, T., and Harrigan, P. (2007b). Predicting HIV co-receptor usage based on genetic and clinical covariates. *Antivir Ther*, page (in press).

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941.

Sirois, S., Sing, T., and Chou, K.-C. (2005). HIV-1 gp120 V3 loop for structure-based drug design. *Curr Protein Pept Sci*, 6(5):413–422.

Skrabal, K., Low, A. J., Dong, W., Sing, T., Cheung, P. K., Mammano, F., and Harrigan, P. R. (2007). Determining human immunodeficiency virus coreceptor use in a clinical setting: degree of correlation between two phenotypic assays and a bioinformatic model. *J Clin Microbiol*, 45(2):279–284.

Svicher, V., Sing, T., Ceccherini-Silberstein, F., Artese, A., Santoro, M. M., Gori, C., Bertoli, A., Alcaro, S., Palamara, G., d'Arminio Monforte, A., Antinori, A., Lengauer, T., and Perno, C. F. (2007). Evolutionary dynamics of NRTI resistance mutations during interruption of anti-HIV therapy. Submitted. **Article appears with note: V.S. and T.S. contributed equally to this work.**

Svicher, V., Sing, T., Santoro, M. M., Forbici, F., Rodríguez-Barrios, F., Bertoli, A., Beerenwinkel, N., Bellocchi, M. C., Gago, F., d'Arminio Monforte, A., Antinori, A., Lengauer, T., Ceccherini-Silberstein, F., and Perno, C. F. (2006). Involvement of novel human immunodeficiency virus type 1 reverse transcriptase mutations in the regulation of resistance to nucleoside inhibitors. *J Virol*, 80(14):7186–7198. **V.S. and T.S. contributed equally to this work.**

Verheyen, J., Litau, E., Sing, T., Däumer, M., Balduin, M., Oette, M., Fätkenheuer, G., Rockstroh, J. K., Schuldenzucker, U., Hoffmann, D., Pfister, H., and Kaiser, R. (2006). Compensatory mutations at the HIV cleavage sites p7/p1 and p1/p6-gag in therapy-naive and therapy-experienced patients. *Antivir Ther*, 11(7):879–887.

## Book chapter

Sing, T. and Däumer, M. (2006). Interpretation algorithms. In Geretti, A. M., editor, *Antiretroviral resistance in clinical practice*, chapter 6, pages 43–56. Mediscript, London.

## Award-winning conference abstracts/posters

Ceccherini-Silberstein, F., Svicher, V., Sing, T., Santoro, M., Beerenwinkel, N., Gago, F., Bertoli, A., Forbici, F., Narciso, P., d'Arminio Monforte, A., Antinori, A., and Perno, C. (2005). Involvement of novel HIV-1 reverse transcriptase mutations in the highly ordered regulation of NRTI resistance. In *14th Int. HIV Drug Resistance Workshop*, Quebec, Canada. Abstract 96. **Best poster award.**

Sander, O., Sing, T., Sommer, I., Low, A. J., Cheung, P. K., Harrigan, R. P., Lengauer, T., and Domingues, F. S. (2007). Prediction of HIV-1 coreceptor usage based on structural descriptors of the gp120 V3 loop. In *Proc. 5th European HIV Drug Resistance Workshop*, Lisbon, Portugal. Abstract 66. **Best poster award.**

Sing, T., Beerenwinkel, N., Kaiser, R., Hoffmann, D., Däumer, M., and Lengauer, T. (2005). Geno2pheno[coreceptor]: a tool for predicting coreceptor usage from genotype and for monitoring coreceptor-associated sequence alterations. In *Proc. 3rd European HIV Drug Resistance Workshop*, Athens, Greece. Abstract 96. **Best poster award.**

Svicher, V., Ceccherini-Silberstein, F., T.Sing, Santoro, M., Beerenwinkel, N., Rodriguez, F., Forbici, F., d'Arminio Monforte, A., Antinori, A., and Perno, C. (2005). Additional mutations in hiv-1 reverse transcriptase are involved in the highly ordered regulation of nrti resistance. In *Proc. 3rd European HIV Drug Resistance Workshop*, Athens, Greece. Abstract 63. **Best oral presentation award.**